ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 22-510

Does Monitoring Change Teacher Pedagogy and Student Outcomes?

Aaron Phipps
United States Military Academy, West Point

Using administrative data from D.C. Public Schools, I use exogenous variation in the presence and intensity of teacher monitoring to show it significantly improves student test scores and reduces suspensions. Uniquely, my setting allows me to separately identify the effect of pre-evaluation monitoring from post-evaluation feedback. Monitoring's effect is strongest among teachers with a large incentive to increase student test scores. As tests approach, unmonitored teachers sacrifice higher-level learning, classroom management, and student engagement, even though these pedagogical tasks are among the most effective. One possible explanation is teachers ``teach to the test'' as a risk mitigation strategy, even if it is less effective on average. This is supported by showing teaching to the test has a smaller effect on student test score variance than other teaching approaches. These results illustrate the importance of monitoring in contexts where teachers have the strongest incentive to deviate from pedagogically sound practices.

VERSION: January 2022

Suggested citation: Phipps, Aaron. (2022). Does Monitoring Change Teacher Pedagogy and Student Outcomes?. (EdWorkingPaper: 22-510). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/7021-1x97

Does Monitoring Change Teacher Pedagogy and Student Outcomes?

Aaron Phipps*

Abstract

Using administrative data from D.C. Public Schools, I use exogenous variation in the presence and intensity of teacher monitoring to show it significantly improves student test scores and reduces suspensions. Uniquely, my setting allows me to separately identify the effect of pre-evaluation monitoring from post-evaluation feedback. Monitoring's effect is strongest among teachers with a large incentive to increase student test scores. As tests approach, unmonitored teachers sacrifice higher-level learning, classroom management, and student engagement, even though these pedagogical tasks are among the most effective. One possible explanation is teachers "teach to the test" as a risk mitigation strategy, even if it is less effective on average. This is supported by showing teaching to the test has a smaller effect on student test score variance than other teaching approaches. These results illustrate the importance of monitoring in contexts where teachers have the strongest incentive to deviate from pedagogically sound practices.

JEL: I21, I28, H75, J24, J41, J45
Keywords: Education Policy, Education Quality, Public School Teachers, Labor Contracts
Approximate Word Count: 12,000

*Assistant Professor of Economics. United States Military Academy, Department of Social Sciences, 607 Cullum Road, West Point, NY 10996. Email: aaron.phipps@westpoint.edu Phone: (845) 549-4697. Special thanks to Sarah Turner, William Johnson, James Wyckoff, and Leora Friedberg for helpful comments and direction. Thanks are also due to the many participants in conferences and other presentations or conversations. The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant #R305B140026 to the Rectors and Visitors of the University of Virginia. The opinions expressed are those of the author and do not represent views of the institute or the U.S. Department of Education, the U.S. Department of Defense, the US Army, or the United States Military Academy.

I. Introduction

Attracting, retaining, and motivating good teachers is an ongoing issue that disproportionately affects students from low-income and minority homes (Lankford et al., 2002; Scafidi et al., 2007; Jackson, 2009). Good teachers measurably improve student scores on standardized testing and life outcomes ranging from decreasing the likelihood of teen pregnancy to increasing college attendance and lifelong earnings (Rockoff, 2004; Rivkin et al., 2005; Aaronson et al., 2007; Kane and Staiger, 2008; Chetty et al., 2014). Yet traditional programs to improve teacher quality, such as teacher training and increased teacher education, have not been sufficient (Hanushek, 2007; Weisberg et al., 2009). As an alternative approach, D.C. Public Schools (DCPS) introduced an unprecedented teacher performance-based incentive program called IMPACT in the 2009-10 school year. In grades with standardized testing, teachers can receive large bonuses by improving their students' test scores and by performing well on unannounced in-class evaluations ("monitoring").

The unannounced evaluations create a unique natural experiment since teachers are effectively drawn at random without replacement. This changes the number of teachers who could be drawn next and – subsequently – the probability of being drawn. In addition to varying the probability of an evaluation, the strict time-frames in which evaluations must occur means teachers may also experience unexpected pockets of unmonitored time. I use both forms of variation to identify how monitoring affects teacher behavior and student outcomes. I find monitoring improves student outcomes by preventing teachers from "teaching to the test" when they are most inclined to do so. Monitoring encourages better classroom management, pacing, and developing higher-level learning. These practices increase student performance on standardized tests, reduce student disciplinary issues, and reduce the daily likelihood that a teacher suspends a student.

The DCPS context provides a unique opportunity to measure in detail how teachers change their behavior. I compare teachers facing intense monitoring (high probability of an evaluation) to teachers facing low monitoring (low probability of an evaluation) to tease out how teachers change their behavior. The in-class evaluations measure teacher performance on nine teaching standards, providing a rich set of measures. It is unsurprising that in the period leading up to standardized testing, teachers clearly shift to teaching to the test when they are monitored less. Less intuitive, however, is the result that teaching to the test is less effective at improving student scores than better pedagogical approaches. Given the large test-based incentive in DCPS, it is not immediately obvious why teachers would choose to use less effective teaching practices *only* in the months before standardized tests. One possible explanation that has received relatively little

attention is teaching to the test may be a risk-mitigating approach (Phipps, 2020). The marginal effect of teaching to the test varies less than pedagogically sound teaching styles, even if the average effect is smaller.[1] I find preliminary evidence consistent with this hypothesis by showing teaching-to-the-test styles reduce student score dispersion.

I add to the literature by showing monitoring changes behavior *and* student outcomes, and then by identifying patterns in how (and when) teachers change their behavior. Monitoring has been shown to work in other contexts, particularly in curbing more egregious behavior. Research in laboratory experiments consistently find monitoring to be the most effective motivator over other forms of incentives (Nalbantian and Schotter, 1997; Camerer and Weber, 2012). In field experiments, researchers found monitoring was particularly effective at preventing unwanted behavior in serial offenders at a charity call center (Nagin et al., 2002). However, evidence on how monitoring affects teachers is more limited. In developing economies, monitoring appears to improve teaching. Duflo et al. (2012) find that using cameras to monitor teacher attendance in India reduced absenteeism by 21 percent and increased student test scores. In the U.S., Dee and Keys (2004) show an incentive program based on unannounced evaluations improved test scores, but their setting does not allow them to separate the motivational effects of monitoring from the learning effects of receiving feedback. Moreover, their causal interpretation is limited given teachers self-selected into the program. Given the mixed results on the effects of in-class observation feedback, distinguishing the two mechanisms is critical for policy design (Taylor and Tyler, 2012; Dee and Wyckoff, 2015; Stecher et al., 2016; Bleiberg et al., 2021). More recently, Phipps and Wiseman (2021) show monitoring changes teacher behavior as measured by evaluation scores, but the authors cannot determine if these performative improvements affect student outcomes in their setting.

This paper also adds to the theoretical discussion on incentives. In general, output-based incentives are preferable to imperfect monitoring because they take advantage of an agent's information advantage (Lazear, 1986; Prendergast, 1999; Neal, 2011). However, incentives based on student test scores have had mixed effects on those same scores, except when paired with monitoring (Pham et al., 2020). I can begin to explore how teachers respond to high-stakes testing by measuring changes in their behavior as standardized tests approach. Intuitively, we would expect monitoring to be less necessary (and less effective) as standardized tests approach, yet I find the opposite. My results shed some light on why test-based incentives are most consistently effective when paired with in-class observations: high-stakes standardized tests

[1] Some researchers find teaching to the test is no more effective – or even less effective – at improving test scores than traditional methods (Blazar and Pollard, 2017; Herman and Golan, 1993; Neill, 2003; Beers, 2005)

appear to induce teachers to teach to the test even though on average it is less effective, and monitoring appears to dampen this behavior. The remaining theoretical question is why this might be, which I can only answer speculatively.

Identifying an effective incentive design for teachers is a top priority to policy makers as it addresses all three goals to recruit, motivate, and retain effective teachers (Dee and Wyckoff, 2015). In a recent survey, high school students cited the lack of career progression and low pay as the top two reasons they would not consider becoming teachers (Croft et al., 2018). Hoxby and Leigh (2004) find that salary compression accounts for 80 percent of the decline in teacher aptitude. Well-designed incentives can improve the career trajectory of quality teachers, both encouraging more students to pursue teaching and retaining high-quality teachers. On the other hand, poorly designed professional accountability may have the opposite effect. Based on survey results from the nationally representative School and Staffing Survey, bonuses based on student test performance correlate with decreases in morale and a sense of value. Test-based bonuses also correlate with a decreased sense that teachers are rewarded for a job well done, a decreased sense of autonomy, overall decreases in job satisfaction, and decreases in satisfaction with salaries and pay.[2] The results here help identify in which contexts we should expect monitoring to be effective, which may provide an alternative to test-based incentives.

My results show that monitoring is effective at reducing unwanted teacher behaviors, which is important in contexts where teachers have especially difficult circumstances and may be inclined to resort to ineffective teaching styles. This is consistent with Dobbie et al. (2013). There, the authors provide evidence that frequent classroom visits and feedback is one characteristic of effective inner-city charter schools. Overall these results are consistent with the existing empirical literature on monitoring. Nagin et al. (2002) find that a sizeable portion of employees respond to monitoring intensity by cheating less as the probability of being audited increases. Interestingly, many employees did not respond to monitoring intensity. In their analysis, the authors find that employees who view their employer as uncaring or unfair are the ones most likely to cheat as monitoring decreased. In the teaching context, this implies teachers may be more likely to "cheat" if they view their employer (or performance incentive) as unfair (Jacob and Levitt, 2003; Sass et al., 2015; Martinelli et al., 2018).

[2]Based on author's calculations.

II. Motivation and Research Context

II.A The Theory of Incentive Contracts as Applied to Teachers

School districts historically attempted to improve teacher quality by promoting more teacher training and experience with what is called a "steps and lanes" system. These pay schemes make salaries depend only on education, certification, and teaching experience. The available evidence shows that these factors do not translate into improved student outcomes (Rivkin et al., 2005). Such pay systems are discouraging to young but effective teachers looking to distinguish themselves in their career, and they provide no credible method for acknowledging and celebrating effective teachers. An alternative payment scheme would seek to identify effective teachers based on their performance and reward them accordingly. This payment scheme, in theory, would encourage high-quality teachers to self-select into teaching and provide appropriate rewards for being productive.

The problem of creating optimal incentives for teachers falls within contract theory. The basic intent of incentive contracts is to reduce employee moral hazard given their asymmetric information about their own effort and talents. Broadly speaking, incentive contracts can depend on performance – an outcome-based incentive – or on some imperfect measure of employee effort (monitoring).

Much of the teacher incentive literature focuses on performance pay based on student test scores. Teaching, however, has well-known characteristics that complicate output-based incentive design (Murnane and Cohen, 1986; Dixit, 2002).[3] Teachers are motivated agents, which makes their response to incentives more inelastic (see Dixit, 2002; Francois, 2000, for example). Motivated teachers may prioritize certain outcomes that do not necessarily align with those of the general public (Neal, 2011). Teachers are also responsible for improving a variety of outcomes that are hard to define, making it unclear which outcome should be used for determining bonus payments. On the other hand, rewarding teachers for multiple outcomes is likely inefficient (Holmstrom and Milgrom, 1991). Even assuming standardized tests capture the most important outputs of teaching, should teachers be paid for their individual contributions to student outcomes or for their team's contribution? Teachers often collaborate, which implies individual,

[3] This is covered more completely in the comprehensive contract theory reviews of Lazear (2012) and Prendergast (1999), which also highlight several concerns that are key for my analysis: how well the measured output aligns with the desired outcome (Akerlof and Kranton, 2005; Neal, 2011; Benabou, 2016), potential gaming of the outcome measure (Baker, 1992), and the use of subjective measures of employee output (Levin, 2003; MacLeod, 2003; Gibbs et al., 2004).

rank-based incentives could reduce collaboration, but a bonus for group achievement introduces moral hazard (Holmstrom, 1982; Kandel and Lazear, 1992).[4] And lastly, if teachers are rewarded based on their contribution to student test scores ("value-added"), their response will likely be muted since value-added scores have inherent noise (Lazear and Rosen, 1981).

Rather than using test scores, incentives could be based on measures of teacher effort through unannounced in-class visits (monitoring). Designing effective in-class observations has a large literature of its own; schools have performed such observations for decades. But from an incentive perspective, existing systems are effectively non-binding: these evaluations are largely perfunctory and virtually all teachers receive passing scores (Weisberg et al., 2009). This problem in part prompted considerable research funded by the Bill and Melinda Gates Foundation into improved methods for measuring effective teaching (Kane and Cantrell, 2010; Kane and Staiger, 2012). With progress in measuring effective teaching practices, it is now more feasible to use in-class observations as an incentive device.

Theories on monitoring are relatively straightforward. Early models focused exclusively on the "rational cheater" assumption: Employees behave as rational agents looking to shirk (cheat) so long as the marginal benefit outweighs the marginal cost of being caught (Becker, 1968). In this vein, Lazear (2006) explores aspects of monitoring to consider, though he does so in the context of crime and test taking. He identifies optimal conditions relating to whether the evaluation criteria should be revealed ahead of time, which is particularly relevant in the teaching context if teachers are able to make superficial changes to their practice to boost their evaluation.[5] He shows when the probability of being evaluated is small or the cost of conducting evaluations is high (as it is in the teaching context), principals should reveal the evaluation criteria ahead of time. Nagin et al. (2002) introduce alternative theories of opportunistic behavior outside the rational cheater model. One of these is the Conscience Model in which employees assume identities that are inconsistent with opportunism (see also Akerlof, 1982). In such cases, monitoring can be reduced in favor of developing a culture and identity inconsistent with opportunism. In the teaching context, shirking will be less likely if teachers are conscience-driven and identify strongly with their role as educators. If true, this would imply monitoring is less effective (and less necessary) for teachers. This may be less true when school districts implement test-based incentives that externalize

[4] As evidence of the moral hazard of group incentives, Imberman and Lovenheim (2015) evaluate the effect of grade-based teacher incentives and find that increasing a teacher's portion of the students in a grade leads to higher test outcomes (to a point). In a randomized trial in New York, Fryer (2013) found that a school-level incentive had no effect on any measured outcome.

[5] Phipps and Wiseman (2021) test this possibility and find that teachers do make general performance improvements as the probability of an evaluation rises, but they do not appear to game the system by focusing on specific, easy-to-adjust evaluation criteria.

existing internal motivation (as in Benabou and Tirole, 2003). These speculative issues have received little empirical attention, at least in developed nations.[6]

II.B Teacher Performance Incentive Programs in the U.S. and Their Effects

In other contexts, there is strong evidence that performance incentives improve employee effort. For example, Lazear (2000) finds that a piece-rate wage in the Safelite Glass Corporation led to significant improvements in output. Nagin et al. (2002) evaluate the effect of monitoring employees in a call center who self-report pledges. They find increased monitoring reduces cheating, though the effects are heterogeneous. In a firm-level randomized experiment, Bandiera et al. (2005) find piece-rate incentives increase employee productivity on a farm, and Bandiera et al. (2007) show managers receiving a performance incentive increase the productivity of their fruit-picking team. These results partly motivated policy makers in the U.S. to actively encourage performance incentives with large-scale federal programs like Race to the Top and the Teachers Incentive Fund. School districts responded by implementing teacher performance incentives that vary considerably in their implementation details.

Empirical evidence on the effects of teacher-level performance incentives has been mixed, but there have been distinctly effective programs. Dee and Keys (2004) exploit the random student assignment from the Tennessee STAR experiment, which overlapped with the implementation of the Career Ladder. This program awarded career advancement and bonuses to teachers for achieving milestones in their in-class evaluations, though teachers self-selected into the program. The authors find students of teachers enrolled in the incentive program improved math and reading scores. Their context, however, does not allow them to determine if the effect is due to monitoring or the result of improvements made post-evaluation, a crucial distinction for incentive design. In the DCPS context, Dee and Wyckoff (2015) use a regression discontinuity approach around the sharp cutoffs in the IMPACT program. They find dismissal threat improved a teacher's effect on student test outcomes by up to 0.05 standard deviations when compared to teachers just above the threshold in the previous year. Also in the DCPS context, Phipps and Wiseman (2021) find teachers change their pedagogy as the probability of an evaluation grows. But because they focus on teachers without test-based incentives, there are few ways for them to determine how the observed pedagogical changes affect students. Other work has found some evidence that the Teacher Advancement Program (TAP), which is a comprehensive teacher review and reward system, improved student outcomes, but these results are less robust (Mann

[6]Some examples in other contexts include India (Duflo et al., 2012), Kenya (Glewwe et al., 2010), Tanzania (Mbiti, 2016; Mbiti et al., 2019; Mbiti and Schipper, 2021) and Mexico (Martinelli et al., 2018).

et al., 2013).

Other performance pay programs analyzed by economists did not dictate specifically how a school district was to create teacher incentives. For example, the Minnesota Quality Compensation (Q-Comp) program, started in 2005, only requires school districts implement an incentive program. But the legislation enacted does not specify the incentive structure. Q-Comp had some small positive effects, but it is unclear what incentive design elements lead to these positive effects (Sojourner et al., 2014). As for the results of the many programs funded by the Teacher Incentive Fund (TIF), the Department of Education's Institute of Education Sciences produced a report that shows small positive effects after three years, but like Q-Comp, there is no clear mechanism (Speroni et al., 2020).

There have been other sizable programs that showed little or no effects on student test scores. The Tennessee Project on Incentives in Teaching (POINT) was a three-year experiment started in 2006. While selection into the experiment was voluntary, assignment to treatment was randomized. Treated teachers would receive bonuses based solely on the test score improvements of their students. There were no significant positive effects from the incentive (Springer et al., 2011). The Denver Professional Compensation program (ProComp), started in 2007, created several routes for teachers to receive bonuses, but by far the largest bonus was awarded to teachers with large gains in student test scores. A report from the University of Colorado, Boulder finds that this incentive had no positive effects on student test scores (Briggs et al., 2014).

Performance incentives for teachers can work, but it is unclear what design elements matter. Intuitively, the size of the incentive ought to matter, yet in a survey of TIF programs, Speroni et al. (2020) find that incentive size is not an important factor. In fact, of the incentive programs described above, the incentive sizes are mostly comparable. The ineffective Denver ProComp and POINT programs both had bonuses ranging between $5,000 and $15,000 (in current dollars), and sometimes even more. Yet the effective Career Ladder program, which did not use any test-based measures of teacher effectiveness, had pay increases ranging from roughly $2,000 to $4,000 in current dollars. One aspect that does appear to correlate with an effective incentive program is using unannounced, differentiating in-class evaluations (monitoring), yet existing evidence does not identify the effects of monitoring specifically.[7]

[7]See Pham et al. (2020) for a complete meta-analysis of incentive programs in the U.S. Notable programs with individual-level incentives based on student test scores are studied in Dee and Wyckoff (2015), Dee and Keys (2004), Hudson (2010), Sojourner et al. (2014), Speroni et al. (2020), Atteberry et al. (2015), and Springer et al. (2011). Of these, only the first five include in-class observations as part of the incentive, while the last two do not. Only the first five show significant, positive effects. Notable programs with grade- or school-level bonuses – with varying levels of effectiveness – include the School-wide Bonus Program in New York (Fryer, 2013), the Dallas School Accountability and Incentive Program (DSAIP) (Ladd, 1999), the Kentucky Instructional Results Information System (KIRIS) (Koretz and Barron, 1998), the North Carolina ABC program (Vigdor, 2008), the Chicago version of TAP (Glazerman

The gap this paper fills is to assess the extent to which monitoring can affect teacher behavior and student outcomes. What remains unclear in the literature is whether in-class observations are effective (Bleiberg et al., 2021), and when they are, is it because of monitoring or because of feedback (Kraft and Christian, 2021)? In results that mirror the larger monitoring literature (Nagin et al., 2002; Nalbantian and Schotter, 1997), I find monitoring has the strongest effect on student outcomes when it prevents unwanted teaching approaches. I control for the possibility that teachers are able to implement their feedback immediately and increase their students' performance, but I find little evidence that feedback affects veteran teachers' students, at least in the year in which teachers received the feedback.[8]

III. Data and Empirical Approach

III.A Setting and Data Source

The IMPACT program began in the 2009-10 school year and its structure was unchanged for the first three years. Over this time period, DCPS had between 128 and 133 elementary, middle, and high schools, with roughly 3,500 teachers each year. Of these teachers, about 13 percent (475 each year) teach grades and subjects for which a teacher's value-added score can be calculated.[9]

As part of the IMPACT incentive program, all teachers receive evaluations from both principals and external evaluators – district employees – called "Master Educators." Principals conduct three evaluations throughout the year and master educators conduct only two. Principals are required to inform teachers a day in advance of their first principal evaluation, but the remaining evaluations are unannounced. In the first year of IMPACT (2009-10), master educators also announced their first evaluation but not their second. The in-class observation uses a well-defined observation rubric called the "Teaching and Learning Framework" (TLF). TLF is a 9-dimensional grading rubric derived from the Danielson Framework. For each dimension, teachers receive a score between 1 and 4. The final TLF score is the average of the scores for all 9 dimensions.

A teacher's final IMPACT score is between 100 and 400. For teachers in grades 4 through 8, the IMPACT score assigns 50 percent weight to a teacher's value-added score and 35 or 45

and Seifullah, 2012), and Houston's ASPIRE program (Imberman and Lovenheim, 2015)

[8] Because there is no variation in the number of evaluations teachers receive each year, I cannot determine the between-year effects of receiving feedback, which is where previous studies have focused.

[9] Value-added scores require a teacher's students have a prior test score available. These scores are only available starting from grade 3 through grade 8, and are only available for Math and English Language Arts (ELA). This means that only teachers in grades 4 through 8 in math and reading will have value-added scores available.

percent weight to classroom evaluations, depending on the year. The remaining weight is assigned to a teacher's score on the "Commitment to School and Community" rating determined by the principal. Based on their overall numeric IMPACT score, teachers receive a rating of "Ineffective" (score below 175), "Minimally Effective" (between 175 and 250), "Effective" (between 250 and 350), or "Highly Effective" (greater than 350). Teachers face large consequences based on their IMPACT rating. Highly Effective teachers receive one-off bonuses ranging from $5,000 to $25,000 depending on the school, grade, and subject taught. If teachers are Highly Effective a second year in a row, they receive permanent pay increases that range from $6,000 per year and possibly exceed $20,000 per year.[10] If a teacher is rated Minimally Effective, she experiences a pay freeze, meaning her salary does not increase as it normally would with each year of experience. She must also improve to Effective in the next year or be dismissed. Receiving a rating of Ineffective leads to immediate dismissal. Only 1.6 percent of teachers received a final rating of Ineffective in the years studied, whereas 12 percent received a Minimally Effective overall rating.

I observe individual student test scores in a teacher's class, the date of each of her in-class performance evaluations, and the date on which she meets with her evaluator to review her performance. With this information, I calculate the number of days in which she is guaranteed not to receive an evaluation and the daily probability of receiving an evaluation. I can also estimate the cumulative effect of having an additional day in which to implement feedback from an evaluation.

III.B Description and Calculation of Treatment Measures

In the IMPACT program, evaluations occur in multiple pre-specified time windows. The structure of these windows provides three opportunities for unmonitored time. In-class evaluations must occur within the time frames depicted in Figure 1. The first principal evaluation must occur by December 1, the second must occur before March 15, and the third must occur before the end of the school year. The Master Educator evaluations split the school year: the first occurs before February 1 and the second occurs afterwards.

Once a teacher has received all of her possible evaluations in the current window, it is guaranteed that she will not have an evaluation until the next window begins. Figure 2 provides an example of how unmonitored time is calculated. Because the first principal evaluation is announced before-hand, I do not consider it monitored time (similarly for the first master educator evaluation in 2009-10). The first window occurs after the first Master Educator

[10] Pay increases depend on a variety of factors, such as a teacher's current base pay, whether her school is a high-poverty school (60 percent or more of students receive free or reduced-price lunch), or if she teaches a high need subject. See Dee and Wyckoff (2015) for more details.

evaluation, but only until the start of the second principal evaluation window. The second possible window of unmonitored time starts from the time of the second principal evaluation and lasts until the start of the second Master Educator window. If a teacher does not receive her second principal evaluation before the start of the second Master Educator window, she will not have any unmonitored days. The third window is possible from the end of her second Master Educator evaluation until the start of the window for the third principal evaluation.[11]

An additional feature that provides empirical identification is how monitoring intensity changes throughout the year. Teachers are effectively drawn without replacement, making the daily probability of an evaluation for a specific teacher increase as the school year progresses. I leverage this variation to identify how teachers change their pedagogy when they are monitored less. I can do this by estimating the teacher's probability of being evaluated on each day. Intuitively, if a teacher has not been evaluated by the last day of the window, she can be certain to receive her evaluation on the next day. My data shows the date of each observation for each teacher, which I use to calculate how likely the remaining teachers are to be evaluated in each of the remaining days. Two factors determine a teacher's estimate of the probability of being evaluated on any particular day: the number of teachers that remain to be evaluated and how many evaluations a teacher expects to be conducted at her school. It is then straightforward to calculate evaluation probability if each remaining teacher has an equal probability.

Let v be an evaluation indicator, where v is $P1$, $P2$, or $P3$ for the principal evaluations and $M1$ or $M2$ for master educator evaluations. Then let a teacher's estimate of the number of evaluations to be conducted on day t at school s be $\hat{L}^v_{ts}$. If R^v_{ts} is the number of remaining teachers needing evaluation v, then each remaining teacher's probability of being evaluated is

$$p^v_{ts} = \frac{\hat{L}^v_{ts}}{R^v_{ts}}.$$

It is straightforward to determine the number of remaining teachers for an evaluation, R^v_{ts}. But estimating how many evaluations a teacher expects to be conducted, $\hat{L}^v_{ts}$, requires assumptions about what a teacher knows. If a teacher knew exactly how many evaluations would be conducted on every day, then $\hat{L}^v_{ts} = L^v_{ts}$, which I observe directly. This is a strong assumption that is unlikely to be true, especially if evaluations are not evenly distributed within a window.

Principals tend to cluster their evaluations near the last third of the time window, which

[11] A small fourth window is possible if all evaluations for a teacher are completed before tests. This window is excluded from the analysis. Because principles rarely complete evaluations at the beginning of their window, unmonitored time in this fourth window only occurs 19 times over the three years across the whole sample (3% of the sample) (see Table 6). Of those instances, 60% of their evaluations occur within 1-2 weeks of standardized testing, making identification unreliable.

causes the expected number of daily evaluations to change over time as well. In the beginning of an observation window, teachers expect principals to conduct few evaluations but increase towards the end of the window. On the other hand, master educators distribute their evaluations more evenly, so the expected number of evaluations remains constant. Figure 3 shows the overall distribution of evaluations across each window. While the master educators maintain a fairly uniform distribution, principals are very often conducting evaluations in the last third of the available time. The dip in evaluations in M2 around day 45 is a result of student testing days in April.

Instead of assuming teachers know exactly how many evaluations will be conducted on each day, I can allow a teacher to assume a uniform distribution of evaluations or assume she is broadly aware of the trend in evaluations. I approximate the information available to a teacher by estimating the distribution of evaluations with a kernel density. The kernel smoothing approximates changes in the trend of daily evaluations that teachers notice. I estimate the model under both a uniform assumption and the kernel, and it turns out not to matter.

I estimate the effects of evaluation probability separately for principal evaluations and Master Educator evaluations. For many days in the year, a teacher has the possibility of either a principal evaluation or a master educator evaluation (or both). The two events are independent and in rare cases both occur on the same day for a single teacher. I use the probability that any evaluation will occur on a specific day.

III.C Empirical Approach

My empirical approach can be built up from a basic model of monitoring. A teacher chooses her vector of effort, x, across her possible inputs (in the DCPS context, there are 9 measured dimensions). She has a baseline choice of inputs x' when she is unmonitored. She receives $\sum b_i x_i$ as a bonus, where b is a vector of pay weights assigned to each of the possible inputs in x. She has costs for each input i which are equal to $c_i x_i + \frac{1}{2} d_i x_i^2$. On monitored days, the teacher maximizes her expected utility:

$$\max_x \; p \sum b_i x_i - \sum \left[c_i x_i - \frac{1}{2} d_i x_i^2 \right]$$

where p is the probability of being observed. The first-order condition for each input x_i is $pb_i - c_i - d_i x_i = 0$. Let x'' indicate her effort on monitored days, then her utility-maximizing choice for input i is $x_i'' = \frac{pb_i - c_i}{d_i}$ (assuming no corner solutions). The principal can encourage and discourage particular practices by her choice of b_i and the intensity of monitoring, p, which provides the mathematical justification for using of monitoring intensity as a treatment.

For estimation, I assume a student's test scores are cumulatively affected by the daily

teaching decisions of the teacher. Given N total instruction days, with m of them monitored and n unmonitored, the teacher contributes $y(x)$ to her student's test score each day. The teacher's total contribution to a student's test score is simply $Y = ny(x') + my(x'')$.

On some days, particularly towards the end of an evaluation window, the daily probability, p, of an evaluation increases. This means $y(x'')$ could potentially change from one day to the next as the probability fluctuates. I can potentially measure these changes in output if I assume the relationship between monitoring probability and output is linear: $y(x'', p) = \bar{y} + \eta p$. This means that if the probability of an evaluation is extremely small, the possibility alone would still induce productivity $\bar{y}$. The variable η is the marginal increase in *output* (due to changes in effort) resulting from p. Substituting into the previous equation, I have

Lastly, teachers receive feedback that they apply to their teaching. Then her contribution on unmonitored days potentially changes: $y(x') = y_0 + \alpha$, where α is the additional daily productivity from receiving feedback. Her feedback is received $\Delta > 0$ days after her evaluation. With this addition, her value-added is

$$Y = \underbrace{(N-n)\bar{y} + \eta \sum_{t=1}^{m} p_t}_{\text{Monitored Days}} + \underbrace{ny_0 + \alpha \sum_{t=m+\Delta}^{N}}_{\text{Unmonitored Days}}$$

The marginal effect of an additional unmonitored day is $y_0 - (\bar{y} + \eta p_m) + \alpha \times 1_{(t \geq m+\Delta)}$. Given the linear structure assumed, this is conveniently rearranged:

$$Y = \underbrace{N\bar{y}}_{\text{constant, } \delta} + \underbrace{\Big(y_0 - \bar{y}\Big)}_{\beta} n + \eta p_m + \eta \sum_{t=1}^{m-1} p_t + \underbrace{(N - m - \Delta)}_{\text{post-feedback days, } q} \alpha$$

I can rewrite this with the substitution δ as the teacher's constant effect, β as the marginal effect of jumping from monitored to unmonitored regardless of monitoring intensity, and q as the number of post-feedback days:

$$Y = \delta + \beta n + \eta \sum_{t=1}^{m} p_t + \alpha q \tag{1}$$

which provides the foundation for using three key variables in my specifications: unmonitored days n, cumulative monitoring intensity $\sum p_t$, and the cumulative feedback days for each evaluation. Then β is the marginal step-wise difference between unmonitored and monitored; η is the marginal productivity of an additional p monitoring intensity for a day, and α is the marginal effect of applying feedback one more day.

III.D Econometric Specifications

My econometric specification builds directly off Equation 1. I use the student test scores directly, Y_{kijs}, for student k with teacher i in year j at school s. Let $w = 1, 2, 3$ indicate the unmonitored window. The number of unmonitored days in window w is n_{ij}^{w}. The number of days between when a teacher receives feedback on evaluation v and student tests is q_{ij}^{v} for $v \in \{P1, P2, P3, M1, M2\}$. I estimate the following equation:

$$Y_{kijs} = W_{kj}\Omega + X_{ij}\Gamma + \phi_s + \delta_i + \sum_{w=1}^{3} \beta^w n_{ij}^{w} + \sum_{v} \alpha^v q_{ij}^{v} + \varepsilon_{kijs} \tag{2}$$

I control for school-level characteristics using school fixed effects ϕ_s and teacher fixed-effects δ_i. Teacher fixed-effects are identified because I observe most teachers multiple years. The variable X_{ij} is a vector of annual teacher experience dummies up to 15 years of experience, teacher race, and pay-scale level. The variables in W_{kj} are student-specific characteristics: previous scores, free-reduced price lunch status, English Language Learner status, special education status, race, and gender. As part of the student controls, all specifications also control for how long the student was in a given teacher's class as a share of the whole year.

I assume that ε_{kijs} is conditionally independent of n_{ijs}^{w}. That is, $E[n_{ij}^{w}\varepsilon_{kijs}|X_{ij}, W_{kj}, \phi_s] = 0$. This amounts to assuming there are no unobservable characteristics of a teacher or student that are correlated with the teacher's contribution to test scores that also systematically change her number of unmonitored days. If evaluators systematically target low-quality teachers or under-performing students early in the year based on criteria that I cannot observe, then my results will be negatively biased.

In order to separately identify the positive effects of evaluation feedback from the effects of unmonitored time, I also assume that the space between when a teacher receives her evaluation and the day she receives feedback is conditionally independent of ε_{ijs}. Keeping the notation from Equation 1, let Δ_{ij}^{ν} be the number of business days between when a teacher received evaluation ν and when she received her feedback. Then in order to interpret α^{ν} as the causal effect of an additional day post-feedback, I am assuming $E[\Delta_{ij}^{\nu}\varepsilon_{ijs}|X_{ij}, \phi_s] = 0$. This is assuming evaluators do not systematically change how long they wait to meet with teachers based on unobservable characteristics that correlate with teacher value-added or student characteristics. If evaluators meet sooner with good teachers, my estimated effects of receiving feedback will be biased upwards.

I estimate Equation 2 using ordinary least squares with clustered errors at the teacher-by-year level. I cluster at the teacher-by-year level because each year at each school is effectively a new

random assignment to treatment, but all students in a class receive the same treatment.[12]

I also look at how teachers change their teaching behavior as the intensity of monitoring changes. I do so by looking at their second Master Educator ($M2$) evaluation scores across the 9 standards. $M2$ is split before and after standardized testing, providing a stark contrast in how teachers change their pedagogy with decreased monitoring both when a standardized test is imminent and when it is not. For a teacher i and standard S on the M2 evaluation at school s and year j, my estimation specification is

$$S_{ijs}^{M2} = \boldsymbol{X}_{ij}\Gamma - p_{ij}\mu + \bar{p}_{ij}\bar{\mu} + \sum_{\nu=P1,M1,P2} S_{ij}^{\nu}\kappa^{\nu} + \boldsymbol{T}_{ij}\omega + \phi_s + \delta_j + \varepsilon_{ijs} \tag{3}$$

where $\boldsymbol{X}_{ij}$ is a vector of experience indicators, ϕ_s is a school fixed-effect and δ_j is a year fixed-effect. As shown in Phipps and Wiseman (2021), the order in which an evaluation occurs matters as well, which is why I include the term $\boldsymbol{T}_{ij}$, a vector of indicators for if the $M2$ evaluation was third (before $P2$), fourth (after $P2$), or fifth (after $P3$). S_{ij}^{q} are scores on standard S for $\nu = P1, M1, P2$, which are evaluations that must occur before testing. The coefficient κ^{ν} captures the correlation between a teacher's performance on standard S in evaluation ν and her performance on that standard during the $M2$ evaluation. The term $p_{ij} = Pr(\text{Eval})$ is measured as the probability of receiving any evaluation on the day of the $M2$ evaluation, and μ is the coefficient of interest. The term $\bar{p}_{ij} = \sum_{d=1}^{D-1} Pr(\text{Eval})_{ijd}$ is the sum of the daily probability of an evaluation and D indicates the day of the evaluation. This last term captures any cumulative preparation effects, which will be important for the instrumental variables approaches that follow.

Given exogenous variation in the probability of an evaluation, I can identify how each teaching standard affects student scores and student score dispersion. I use the two-step efficient generalized method of moments estimator (GMM), where the key variables in Equation 3 operate as instruments.[13] For this analysis, I limit the sample to teachers who receive their M2 evaluation before standardized tests. Let $\boldsymbol{S}_{ijs}^{M2}$ be a vector of the endogenous teacher evaluation scores on each standard in her M2 evaluation. The terms p_{ij}, $\boldsymbol{S}_{ijs}^{\text{P1}}$, $\boldsymbol{S}_{ijs}^{\text{M1}}$, $\boldsymbol{S}_{ijs}^{\text{P2}}$ and $\boldsymbol{T}_{ij}$ identify $\boldsymbol{S}_{ijs}^{M2}$. Additionally, I assume that the cumulative evaluation probability only affects student test scores through a teacher's performance of the teaching standards, making the cumulative probability $\bar{p}_{ij}$ an additional instrument.

[12] Most specifications use the built-in Stata function `areg`. In cases with more fixed-effect estimations, I use the function `reghdfe` (see Correia, 2014).

[13] For all instrumental variable analyses, I use that Stata function `ivreg2` (see Baum et al., 2010).

The specification for measuring the causal effect of standard S_{ijs} on student score Y_{kijs} is:

$$Y_{kijs} = W_{kj}\Omega + X_{ij}\Gamma + \hat{\boldsymbol{S}}_{ijs}^{M2}\eta + \sum_{w=1}^{3}\beta^{w}n_{ij}^{w} + \phi_s + \delta_j + \varepsilon_{kijs} \tag{4}$$

where $Z = \left[\boldsymbol{S}_{ijs}^{\mathrm{P1}}, \boldsymbol{S}_{ijs}^{\mathrm{M1}}, \boldsymbol{S}_{ijs}^{\mathrm{P2}}, \boldsymbol{T}_{ij}, p_{ij}, \bar{p}_{ij}\right]$ are instruments for $\hat{\boldsymbol{S}}_{ijs}^{M2}$.

For measuring the effect of each standard on score dispersion, let D_{ijs} be a measure of student score dispersion for teacher i's class in year j at school s. The estimated model is

$$D_{ijs} = X_{ij}\Gamma + D_{ijs}^{\mathrm{prev}}\rho + \hat{\boldsymbol{S}}_{iys}^{M2}\eta + \sum_{w=1}^{3}\beta^{w}n_{ij}^{w} + \delta_j + \varepsilon_{ijs} \tag{5}$$

The term D_{ijs}^{prev} is the same measure of dispersion for a teacher's current students but calculated using their previous scores from the year before. The terms n_{ij}^{w} are, as before, the number of unmonitored days for each window w.

III.E Data Summary

My sample is limited to students in the fourth and fifth grades due to data limitations. DCPS only recorded teacher-student combinations starting in fourth grade since this is the first grade value-added factors into a teacher's IMPACT score. The school district maintained carefully constructed digital records of which students were in each teacher's class for grades that need to calculate a value-added score. Because teachers in high school (grades 9-12) are not assessed based on student test scores, detailed records connecting students to teachers are not as complete and reliable for high school. Similarly, student assignments are complicated for grades six through eight because of different class structures across schools and grades: some topics are split while other classes are shared, but these distinctions are not observable to me, making grades six through eight unavailable for this analysis.

To be included, students must have a test score available from the prior year, which excludes students in their first year at the district. This is mainly because the preferred specification uses prior student test scores as a control on student ability, though the results are robust across other specifications with the full sample of students. Only students who have been in their teacher's class for more than a quarter of the school year are included in that teacher's class.[14]

I also restricted the analysis to teachers who are not in their first year of teaching. The first

[14]The results are robust to this decision. The results only begin to change when the sample includes students who were in a teacher's classroom for less than 5% of the school year. About 7% of students switch teachers within the school year. In these cases, students were assigned to the teacher they had for the longest period of time.

year of teaching is considered a training period, making responses to monitoring and feedback different from other teachers. Because of their lack of established lesson plans and classroom management skills, first-year teachers are unlikely to have a large skillset – or pedagogical approaches – from which to choose. This is confirmed in a heterogeneous effects analysis on first-year teachers using an interaction term of first-year status with unmonitored days (see Tables A25 and A26 in the online appendix).

With these restrictions, there are about 4,000 students per year in my sample, as shown in Table 1. The student population is 70 percent Black, 15 percent Hispanic, and 11 percent White, which can be seen in the student demographics table, Table 2. This composition remains constant across the years. About seven percent of students are English Learners and 12 percent are enrolled in special education programs. This district and sample represents a fairly low socio-economic status student population, where nearly 70 percent of students receive free or reduced-price lunch each year.

My sample has about 220 teachers each year. The average class size is 18 students, though this only includes students that are in my sample. The average teacher has 11.25 years of experience, which remains constant across the study period (see Table 3). Demographically, the teacher population proportionally matches the Black student population, but White teachers are relatively over-represented and Hispanic teachers are under-represented. The teacher population is 64 percent Black, 2 percent Hispanic, and 31 percent White (Table 4).

There are about 81 schools each year throughout my observation period. School-level demographics remain constant across the study period, but there is meaningfully large variation between schools in student race and the fraction of students with free and reduced-price lunch. The school district is fairly segregated along student race and socio-economic status dimensions (Table A1, Online Appendix). Both teacher racial composition and average teacher experience vary considerably between schools, even after excluding first-year teachers (Table A2, Online Appendix). Taken together, these facts reinforce the need for school-level fixed effects in the econometric specification.

Given the complexity of the context, there are several treatment variables to summarize. The key treatment is unmonitored days. Table 5 reports the average number of unmonitored days in each window *among teachers who experience unmonitored days*. Because of the overlapping evaluation time frames for principal evaluations and external evaluators, many teachers do not have unmonitored days. For Window 1 in 2010, all teachers have exactly 46 school days that are unmonitored since both the first principal and external evaluations were announced at least a few days in advance. When this restriction was lifted in subsequent years, it appears that the first

external evaluation in 2011 occurred later than in 2012. Across all three windows, the average number of unmonitored days varies from year to year, particularly Window 3 with an average of 6 unmonitored days in 2010 and 15 in 2012. Some of this variation can be explained by changes in how the window cut-off dates landed in the week. Evaluations were less likely to occur on Fridays. If the cutoff fell on a Monday or Tuesday, many principal evaluations would often be conducted a full week earlier than if the cutoff occurred on a Thursday or Friday. Another part of the wide treatment variation between years may be explained by principals learning the value of completing their evaluations sooner rather than later.

Table 6 shows the fraction of teachers with unmonitored days for each window. Window 4 is also included to show that it is possible but very unlikely that a teacher receives her last evaluation before her students take their standardized tests. The key take-away is that roughly a fifth of teachers experience unmonitored days in Window 2 and a quarter experience unmonitored days in Window 3. The treatment is unbalanced, which will affect statistical power.

Additional measures used in the analysis include the cumulative probability of an evaluation (shown in Table 7). Average cumulative evaluation probability can exceed 1.0 because it is the sum of daily probabilities. The number of days between receiving evaluation feedback and student test-taking is shown in Table 8. The number of feedback days for the first principal and external evaluator observation are understandably large, which will matter when interpreting the effect size. Finally, there are very few post-feedback days for the last principal evaluation (which usually occurs after standardized testing), as shown in Table 9.

IV. Balance Checks

The justification for interpreting my results as causal rests on the intent communicated to evaluators, conversations with principles and teachers about evaluation procedures, and anecdotal evidence that evaluations were not timed or targeted at specific teachers or students. The measured effect of unmonitored time is causal under the identifying assumption that the number of unmonitored days is independent of unobserved qualities that affect a student's test scores, including the qualities of her teacher:

$$E[n_{ijs}^{w}\varepsilon_{ijs}|X_{ij},\, \phi_s] = 0$$

for student i in year j and school s with $w = 1, 2, 3$. This is likely to be true when neither principals nor external observers choose the timing of their visits based on some quality I do not observe in the data. In practice, external observers would be assigned a school and work through

the teachers at that school. Anecdotally, principals conducted their evaluations on an as-possible and ad hoc basis. The intent communicated to principals and external observers was that they were to be unpredictable, avoiding systematic targeting of a grade or hallway. However, even if unpredictable to the teacher, if observers systematically targeted teachers with the worst-behaving or struggling students, the identification assumption is violated.

Short of an explicit randomization process, the identification assumption is not verifiable. However, I use a variety of balance checks as supporting evidence that the identifying assumption is valid. While the information available to me is not identical to that of the principal or external observer, I have information on both students and teachers about prior performance that I use to check for correlations in treatment and notable characteristics. For students, the observable characteristics I use to check for treatment targeting are gender, race and ethnicity, if they are enrolled in an English Language Learner (ELL) program, whether a student is enrolled in a special education program, if they receive free or reduced-price lunch, and their prior-year math and reading score on standardized tests. In all the checks, coefficients are not adjusted for multiple hypothesis testing. In all the tests conducted, treatment appears balanced.

Balance Check 1: Treated vs Untreated by Student Characteristics

Whether or not a student's teacher experiences any unmonitored days does not correlate with observed student characteristics. I check treated and untreated population composition the Online Appendix Table A3. Within a school, an untreated student was 0.25% less likely to be a boy for Window 1 relative to treated students, which is not statistically significant. The only mildly significant difference (at the 10 percent level) is for English-Language Learners in Window 1 and Hispanic students in Window 2.

Balance Check 2: Number of Unmonitored Days by Student Characteristics

I also find no evidence that the number of unmonitored days correlates with student characteristics. These results can be found in the Online Appendix in Tables A4 -A6. These specifications add teacher experience. They also include fixed-effects for the school, the year, and the subject. There are no consistent patterns in sign or magnitude across windows for previous scores, free and reduced-price lunch, English learner, or special education. The only statistically significant coefficient is for previous math scores, but the effect is positive. The bias would go in the opposite direction of my results. The F-statistics are not significant.

Balance Check 3: Treated vs Untreated by Teacher Characteristics

Along the extensive margin (treated or untreated), evaluators do not appear to target particular teachers. Table A7 in the Online Appendix reports the results of checks for systematic differences in whether a teacher has any unmonitored days based on observable characteristics. The observable characteristics considered are a teacher's prior experience, her race and ethnicity, her current salary step – which is a proxy for experience and education – and lastly her previous years' evaluation score. None of these characteristics systematically predict a difference in a teacher's likelihood of experiencing any unmonitored time in any of the three windows.

Balance Check 4: Number of Unmonitored Days by Teacher Characteristic

The number of unmonitored days does not consistently correlate with any observable characteristics. Tables A8 -A10 in the Online Appendix show the results of a regression-based balance check with school and year fixed-effects. Additional characteristics in these regressions are whether or not a teacher was ranked Minimally Effective or Highly Effective in the previous year (relative to teachers with an Effective rating). It appears as though teachers that were previously Highly Effective have less unmonitored time in Window 1, though no other variables are significant for any other Windows. The F-statistics are never significant.

Balance Check 5: Time between Evaluation and Feedback by Teacher Characteristics

Causally identifying the effect of feedback on student outcomes rests on an additional assumption: the length of time between an observation and when the evaluator meets with a teacher to discuss her evaluation. If evaluators systematically provide feedback much faster to low-performing teachers, the effects of feedback will be biased down. These balance checks are reported in Tables A11 - A14 of the Online Appendix. A higher lagged evaluation score appears to reduce the amount of time between evaluation and conference for $P1$, though the sign switches for $P2$. If true, the negative coefficient would bias my estimate of the effect of feedback upward. With $P2$, being Minimally Effective in the previous year increases the length between evaluation and conference, which would bias my estimate of the effect of feedback upward as well. The $M2$ evaluation appears to correlate with teacher race, thought it's not clear how this would affect estimates.

V. Results

V.A How Does Unmonitored Time Affect Student Outcomes?

Effects on Student Test Scores

My key result is that for both math and reading, monitoring teachers in the time leading up to standardized tests improves test performance. The results are shown in Table 10 for math outcomes and Table 11 for reading. The 95 percent confidence interval for the p-values of 1,000 randomization inference trials are also shown in brackets for the key coefficients.[15] Each column represents a new specification. The first is the simplest specification with only unmonitored days and teacher, school, year, and grade fixed-effects and previous student test scores. Column two adds the post-evaluation feedback controls and column three adds monitoring intensity controls. The final columns add teacher experience and student demographic information: student race and gender, special education, English language learner, and free or reduced-price lunch status.

Unmonitored days in Window 1 do not appear to have a significant effect on student test outcomes while Window 3 consistently does. Unmonitored time in Window 2 has a significant effect on math scores once I control for feedback time, but not in reading. From row three of the math results, an additional unmonitored day negatively affects student math outcomes from between 0.011 and 0.013 standard deviations. The average teacher has 8.3 unmonitored days in Window 3 if she has any, which is slightly less than two business weeks. The average effect of unmonitored days for such a teacher ranges from -0.11 to -0.13 standard deviations in math for about two weeks of unmonitored time. From the third row of the reading results, the range of average effects is from -0.012 to -0.014 standard deviations and is also statistically significant.

I test the extent to which feedback affects student test scores within the same year. As we would expect, adding controls for feedback time increases the measured effect of unmonitored time by about 0.003 standard deviations for both reading and math, though the difference is not statistically significant. The coefficients for an additional feedback day are not statistically significant but can bee seen in Tables A15 and A16 of the Online Appendix. The feedback coefficients are consistently positive for math, but not consistently statistically significant. For reading, there is more fluctuation in the measured effect of evaluation feedback.

Including the cumulative probability of an evaluation does not change the effect of an

[15] Randomization Inference results calculated using the `ritest` command in Stata (Heß, 2017). Treatment of unmonitored days is randomly assigned at the teacher-year level, preserving the clustered design.

unmonitored day. There are few statistically significant results for the effect of cumulative probability (for the full coefficient results, see Tables A15 and A16 of the Online Appendix). What results there are remain unstable and are not consistent across other specifications. This corresponds to teachers' productivity having a very sharp response to any chance of an evaluation, but less of a response to increasing the probability of an evaluation. Another interpretation is viewing this as evidence of the different effects from deciding which approaches to use (the extensive margin) vs teaching intensity (the intensive margin). While teachers may alter their teaching intensity as monitoring increases through more careful preparation with only mild effects, in the complete absence of monitoring teachers may select different approaches altogether. Lastly, it may be the case that the summation of probabilities is too coarse and teacher effects are not in fact linear.

While post-evaluation time and cumulative evaluation probability are somewhat colinear with unmonitored days, the problem is exacerbated for the first window. Teachers need to only receive their first Master Educator evaluation to have unmonitored time. While the coefficient is technically identified given variation in daily probabilities across schools, the resulting measures of evaluation probability and post-evaluation feedback are strongly correlated with unmonitored time in Window 1. The multicollinearity issue becomes most apparent in the reading results, where the effect of unmonitored days in Window 1 spike *up* when post-feedback days from $M1$ are included, and the effect of post-feedback days is large and *negative*, though the two effects cancel each other out. The other windows do not suffer from such extreme multicollinearity because unmonitored days require both evaluations to be complete, and they can be done in any order. This provides more identifying variation.

My specification assumes the effect of unmonitored days is cumulative and linear. I test this assumption by binning unmonitored days into 5-day increments, shown in Figure 4. The results for Window 3 are robust, though it looks like the effect may have a lag of a couple days. It is also possible that unmonitored time has heterogeneous effects on students. If changes in teaching behavior on unmonitored days disproportionately affect the best-performing students, it would provide some evidence that teachers may be trying to improve the scores of their lowest performers at the expense of their highest performers. I test this possibility with a quantile regression based on student performance quantile, shown in Figure 5. Windows 1 and 2 continue to have no effect on student outcomes. For Window 3, all coefficients are significantly negative, but unmonitored days do not appear to have heterogeneous effects by student quantile. In reading, unmonitored time has a slightly more negative effect on high-performing students, though it is not statistically significant.

Effects on Student Suspensions

There are several other student outcomes beyond test performance that are generally important to teachers, parents, and policy makers. One such outcome would be student disciplinary actions. Intense test preparation and teaching to the test may have adverse effects on student behavior. Sacrificing pacing, classroom management, and student engagement (teaching standard 8) in particular should affect student behavior. Table 12 shows the measured effect of unmonitored time on each teacher's log annual total short-term suspensions (i.e. excludes expulsions and suspensions incurred through criminal activity such as bringing a weapon to school). Suspensions are about 3.2 percentage points higher for teachers with two weeks of unmonitored time (ten days) during Window 3. Notably, unmonitored time in other windows does not appear to affect student discipline.

To determine whether or not these additional suspensions occurred during unmonitored days, Table 13 considers the daily probability of issuing a suspension for monitored and unmonitored days in Window 3. This linear model uses the number of suspensions a teacher issues each day as an outcome. There are about 23 observations per teacher per year. The average teacher issues 0.06 suspensions each day during this time period. As seen in the first row of Table 13, the effect of an unmonitored day increases the rate by between 0.05 and 0.07, which effectively doubles the average number of daily suspensions. These large effects motivate the need to understand more clearly how teachers change their teaching style. In general, we would expect less preparation – and therefore less classroom management – to increase behavioral issues.

Robustness and Sensitivity

I conduct three basic robustness and stability checks. The first is to use the last principal evaluation – which occurs after standardized testing – to check for an effect of unmonitored time on student test scores where there should not be any. The second test is to use the stability testing procedure proposed in Oster (2019) based on the theoretical work in Altonji et al. (2005). Lastly, I evaluate an array of specifications that change the variables used, their definitions, and the sample.

Roughly 97 percent of the final principal evaluations occur after student tests. I find no evidence that student test scores correlate with the placebo treatment, which is unmonitored time occurring *after* students have completed their standardized testing. I use the same specification as before but now include the placebo unmonitored time in Tables 14 and 15. If my main results are spurious, it could be the result of principles targeting their worst teachers (or students) early. Principles might do so in order to help students prepare better for their test, or they might do so

to get harder evaluations done early. The results of the placebo check support my key findings: placebo days have no significant effect on student outcomes.

Using the procedure from Oster (2019) based on the theory in Altonji et al. (2005), I find there would need to be an unlikely degree of correlation between unobservable factors and student test scores to explain my results. The estimation requires assuming a maximum R^2 value if all factors were observable. I construct a table that uses possible R^2 values between my observed R^2 and 1. The results are shown in Table 16. Only under the extreme assumption that the maximum R^2 is greater than 0.95 do I find my results for reading could be explained by unobserved factors if they correlated with the outcome about 90 percent as much as the observable characteristics.

The Online Appendix provides a set of tables testing the sensitivity of my results to various assumptions. In Tables A17 and A18, I show that the results are not sensitive to how I specify experience. I use a continuous measure with a squared term and vary whether or not experience is capped at 20 years. The results remain unchanged.

The main effect of unmonitored days on student test outcomes could potentially be the result of increased absences from suspensions. I test this by dropping all students with any suspensions (Tables A19 and A20, Online Appendix). The sample size drops from 12,305 and 12,820 students for math and reading to 7,416 and 7,787, a 40% drop. The point estimates are slightly smaller for math, though they vary much more. The negative effects of unmonitored in Window 3 drop by about half in reading to around 0.006 standard deviations per unmonitored day. This may indicate that some of the overall effect on reading is caused by absences.

Another assumption in my complete specification is the shape of the distribution of evaluations. Tables A21 and A22, show the results after changing teachers' expectations to believing all evaluations are uniformly distributed. I find there is no meaningful change in my results.

Tables A23 and A24 show results broken out by grade. Interestingly, the math effects are more concentrated in fourth grade than in fifth. The fourth grade math curriculum covers fractions and operations on fractions, while the fifth grade math curriculum covers decimals and their operations. These results suggest attempts to teach fractions without higher-level understanding is less effective than it is for decimals. For reading, there is no such meaningful distinction in the curriculum and there is no observable difference in the effect of unmonitored days for reading between fourth and fifth grade.

Lastly, Tables A25 and A26 include inexperienced teachers with first-year status interacted with the treatment variables. The effect of unmonitored time on inexperienced teachers is large, positive and significant, but switches to negative once I control for feedback. The effects for

first-year teachers are very large, something that may be the result of the volatility of first-year teaching performance. Because of cell sizes, the later columns have large standard errors and lose statistical significance. This is mild evidence that evaluation feedback may matter particularly for first-year teachers.

V.B How do Teachers Change Their Pedagogy When Unmonitored?

What are teachers doing differently as tests approach that would create such differences between monitored and unmonitored time? Without observing teacher activity before and after evaluations, I cannot identify directly what they do when unmonitored. However, variation in the daily probability of an evaluation provides an opportunity to estimate how teachers change their pedagogical approach when they are monitored less intensely, that is, when they are less likely to be evaluated. I do so using teacher scores on their $M2$ evaluation with variation in the probability of being evaluated. The $M2$ evaluation has the advantage of being split before and after standardized testing, allowing me to identify changes in teacher responses to monitoring intensity during and after the intense test-prep season.

Looking at the description of the individual evaluation components (see Table A27 for complete descriptions), there are a few elements that would reflect teachers shifting to a more teaching-to-the-test approach. Standards 3 and 4 would capture teachers focusing too much on some students at the expense of others. These items are labeled "Engage students at all learning levels in accessible and challenging work," and "Provide students multiple ways to move toward mastery." In focusing on the tested content specifically, teachers may fail to provide a variety of learning methods. Standards 6 and 7 are particularly indicative of a teaching-to-the-test mentality, as they touch on probing for and building up a deeper learning. Standard 7 is "Develop higher-level understanding through effective questioning," and Standard 6 captures probing for deeper understanding and building up knowledge gradually ("scaffolding"). Teachers may sacrifice encouraging a deeper, higher-level understanding in order to ensure students grasp tested material. Lastly, Standard 8 measures lesson pacing, student behavior and idleness. All these pedagogical elements could be expected to suffer as teachers are increasingly pressed to ensure their under-performing students meet test requirements. These students are likely to need a slower pacing and have more behavioral concerns that could be exacerbated with overly-intense instruction. At the same time, well-equipped students may lack engagement as teachers focus on rote material.

I measure how teachers change their teaching styles using the specification in Equation 3. The results are in Table 17 and shown graphically in Figure 6 where I've broken the analysis up

between teachers who receive their evaluation before standardized testing and those that received it afterwards. The results show Standards 3, 4, 6, 7, and 8 are where teachers sacrifice the most before standardized tests. The coefficients provide a linear estimate of the difference in teacher behavior when switching from a 100 percent chance of an evaluation to a zero percent chance. When unmonitored, teachers shift their pedagogical priorities rather significantly in the time leading up to standardized tests. On average, these approaches do not appear to pay off given the observed effects of unmonitored time.

It remains to be shown whether or not the sacrificed standards are effective at improving student test scores. The effect of each teaching standard on student test scores (η in Equation 4) is shown in Figure 7.[16] Using multiple learning methods, developing higher-level understanding, pacing, and classroom management (Standards 4, 7, and 8) are particularly effective at improving student test scores both in math and reading.

V.C Is this Risk Mitigation?

It is unclear why teachers would choose to switch away from good teaching practice as tests approach but not at other times. One possible explanation is that teachers are mitigating risk. Given the high-powered, test-based incentives, teachers may be looking to shore up certainty that their students will perform well – possibly at the expense of average performance.

Effects on Score Variation

To test the risk-mitigation hypothesis, I can use a similar IV approach to measure the effect of teaching techniques on student score dispersion. The effect of teaching standards on student test score dispersion (η in Equation 5) are shown in Figure 8. The sample is limited to only teachers who receive their $M2$ evaluation before standardized testing. The results show using multiple learning methods and developing higher-level understanding (Standards 4 and 7) consistently increase student score dispersion. Scaffolding (Standard 6) does so for math scores but not for reading scores. Engaging students at all learning levels (Standard 3) increases dispersion in reading scores but decreases dispersion in math scores. Pacing and classroom management (Standard 8) reduces score dispersion for both reading and math. While somewhat ambiguous, these results verify the intuition that developing higher-level learning and using

[16] In each IV estimation, I test for weak and under-identification for each endogenous regressor ($\boldsymbol{S}_{ijs}^{M2}$) separately using the method in Sanderson and Windmeijer (2016). When measuring each Standard's effect on student test scores, the endogenous variables are well-identified individually. However, in the joint test for under-identification using the rK statistic from Kleibergen and Paap (2006) the model fails to reject the null-hypothesis that the endogenous variables are under-identified. Using the weak identification test (Kleibergen-Paap Wald rK F statistic), the endogenous variables are weakly identified. When measuring the effect of each Standard on student score variation, the individual and full model are all well identified.

multiple learning methods are potentially riskier methods for improving student test scores.

Along these same lines, I can test if teacher behavior during unmonitored times has an overall effect on student score dispersion. I attempt to formally measure this in Tables 18 and 19. These show the change in distance for a teacher's student score distribution between the 95th and 5th, 90th and 10th, 80th and 20th, and the 70th and 30th percentiles. Notably, only in Window 3 is there a reduction in distribution width. Using bootstrapped standard errors, I find that none of these effects are statistically significant.

In all, there is strong evidence that teachers sacrifice specific teaching standards in the time leading up to standardized tests. These changes look like teaching to the test. The standards sacrificed stand out as being effective at improving test scores. There is also evidence that several of these sacrificed standards have particularly varying effects on student outcomes. Together this provides evidence consistent with the hypothesis that teachers are looking to mitigate risk by sacrificing high-average-effect (but high-variance) standards in the time leading up to standardized testing.

What about teachers without a test-based incentive?

If teachers teach to the test to mitigate the financial risks of student test performance, it would be valuable to test the extent to which this behavior is incentive driven. During the 2011-2012 school year, DCPS experimented with using teacher value-added measures for the third grade (though with no incentive attached). As a result, I can link students and teachers for third grade in that school year. I run the same analysis on third grade students, but in this case there are no test-based high-stakes incentives. The results are in Tables 20 and 21. The sample size is greatly reduced, and previous student test scores are not reliably available for second graders from 2010-2011, making it impossible to control for a student's previous test scores. Because the data are for a single year, there are no year fixed-effects or teacher fixed effects. Therefore, the results are not perfectly comparable to the main results reported in Tables 10 and 11, but are most similar to columns 1 through 4 only without teacher fixed effects.

The results show positive effects from unmonitored time in math during Window 3 until I account for evaluation feedback time. In reading, the effect of unmonitored time is smaller than in higher grades and not statistically significant. Moving from column 3 to 4 – which adds student indicators for special education, free and reduced-price lunch, and ELL status – the effect of unmonitored time drops to -0.002. Overall, focusing on Columns 4-6 which include all controls, the effect of unmonitored time in 3rd grade is smaller by nearly half and not statistically significant. While not definitive because of a smaller sample size and the lack of teacher fixed

effects, Tables 20 and 21 are consistent with the idea that much of the observed effects from monitoring are in part driven by the presence of a test-based incentive.

VI. Conclusion

While teachers play an important role in student outcomes, improving teacher quality through policy has proven difficult. Speroni et al. (2020) compare teacher incentive programs along six dimensions ranging from incentive size to teachers' understanding of the performance program. Notably, the use of high-stakes unannounced in-class observations was not a dimension they considered. They conclude that "...none of the characteristics we examined could help explain observed differences in student achievement impacts across districts." If professional accountability is to be a viable policy solution, the research priority is to identify what works.

This paper contributes to that effort by providing causal evidence of how monitoring teachers affects student outcomes. The approach here is unique in that it separately identifies the effect of receiving feedback from the effect of monitoring. In-class evaluations appear to improve teaching, though the effect is limited. For teachers with high-stakes student testing, monitoring helps by preventing unwanted behaviors and less through feedback. Feedback has limited effects – except for first-year teachers – which is consistent with recent studies (Kraft and Christian, 2021; Bleiberg et al., 2021).

The potential downsides of such high-stakes teacher evaluation systems still remain. If poorly designed, teacher evaluations can encourage gaming or over-emphasis on a single component of the evaluation rubric. Teachers may also be reticent to forfeit autonomy over their teaching style, making it pragmatically more difficult to implement. There is a financial cost to implementing evaluations. In 2017, DCPS chose to reduce the number of evaluations due to cost concerns. An "open-doors" policy where principals can briefly observe teaching unannounced may be more cost effective and have the same accountability effects, though more rigorous evidence is needed on this (Dobbie et al., 2013). Evaluating the cost-effectiveness of high- (and low-) stakes in-class evaluations is an excellent avenue for future research. Lastly, it is feasible to make improvements in standardized testing to bolster teacher confidence that the pedagogically sound teaching style is the best approach for preparing their students.

Funding Disclosure Statement for Aaron Phipps

The research reported here was supported by the Institute of Education Sciences, US Department of Education, through Grant #R305B140026 to the Rectors and Visitors of the University of Virginia. The opinions expressed are those of the author and do not represent views of the institute or the US Department of Education.

Administrative data were generously provided by the Washington DC Public Schools without compensation to The Center on Education Policy and Workforce Competitiveness at the University of Virginia. DC Public Schools has been allowed to review the results for factual errors.

IRB approval for this project was obtained through the University of Virginia Institutional Review Board.

References

Aaronson, D., L. Barrow, W. Sander, Daniel Aaronson, Lisa Barrow, and William Sander (2007). Teachers and Student Achievement in the Chicago Public High Schools. *Journal of Labor Economics 25*(1), 95–135.

Adnot, M., T. Dee, V. Katz, and J. Wyckoff (2017). Teacher Turnover, Teacher Quality, and Student Achievement in DCPS. *Educational Evaluation and Policy Analysis 39*(1), 54–76.

Akerlof, G. A. (1982, 11). Labor Contracts as Partial Gift Exchange. *The Quarterly Journal of Economics 97*(4), 543.

Akerlof, G. A. and R. E. Kranton (2005). Identity and the Economics of Organizations. *Journal of Economic Perspectives 19*(1), 9–32.

Altonji, J. G., T. E. Elder, and C. R. Taber (2005, 2). Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools. *Journal of Political Economy 113*(1), 151–184.

Atteberry, A., D. C. Briggs, and S. Lacour (2015). Year 2 Denver ProComp Evaluation Report : Teacher Retention and Variability in Bonus Pay , 2001-02 through 2013-14. Technical report, Colorado Assessment Design Research and Evaluation Center, University of Colorado, Boulder.

Baker, G. P. (1992). Incentive Contracts and Performance Measurement. *Journal of Political Economy 100*(3), 598–614.

Bandiera, O., I. Barankay, and I. Rasul (2005). Social Preferences and the Response To Incentives: Evidence From Personnel Data. *The Quarterly Journal of Economics 120*(3), 917–963.

Bandiera, O., I. Barankay, and I. Rasul (2007). Incentives for Managers and Inequality Among Workers: Evidence From a Firm-level Experiment. *Quarterly Journal of Economics 122*(May), 729–773.

Baum, C. F., M. E. Schaffer, and S. Stillman (2010). ivreg2: Stata module for extended instrumental variables/2SLS, GMM and AC/HAC, LIML and k-class regression.

Becker, G. S. (1968). Crime and Punishment: An Economic Approach. *Journal of Political Economy 76*(2), 169–217.

Beers, G. W. (2005, 7). The Effect of Teaching Method on Objective Test Scores: Problem-based Learning versus Lecture. *Journal of Nursing Education 44*(7), 305–309.

Benabou, R. (2016). Bonus Culture: Competitive Pay, Screening, and Multitasking. *Journal of Political Economy 124*(2), 305–370.

Benabou, R. and J. Tirole (2003). Intrinsic and Extrinsic Motivation. *Review of Economic Studies 70*(3), 489–520.

Blazar, D. and C. Pollard (2017). Does Test Preparation Mean Low-Quality Instruction? *Educational Researcher 20*(10), 1–14.

Bleiberg, J., E. Brunner, E. Harbatkin, M. A. Kraft, and M. Springer (2021, 12). The Effect of Teacher Evaluation on Achievement and Attainment: Evidence from Statewide Reforms.

Briggs, D., E. DiazBbilello, A. Maul, M. Turner, and C. Bibilos (2014). Denver ProComp Evaluation Report: 2010-2012. Technical report, Colorado Assessment Design Research and Evaluation Center, University of Colorado, Boulder.

Camerer, C. F. and R. A. Weber (2012). Experimental Organizational Economics. In R. Gibbons and J. Roberts (Eds.), *The Handbook of Organizational Economics*, Chapter 6, pp. 213–262. Princeton, NJ: Princeton University Press.

Chetty, R., J. Friedman, and J. Rockoff (2014). Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *American Economic Review 104*(9), 2633–2679.

Correia, S. (2014). REGHDFE: Stata module to perform linear or instrumental-variable regression absorbing any number of high-dimensional fixed effects.

Croft, M., G. Guffy, and D. Vitale (2018). Encouraging More High School Students to Consider Teaching. Technical report, American College Testing.

Dee, T. S. and B. J. Keys (2004). Does Merit Pay Reward Good Teachers? Evidence from a Randomized Experiment. *Journal of Policy Analysis and Management 23*(3), 471–488.

Dee, T. S. and J. Wyckoff (2015). Incentives, Selection, and Teacher Performance: Evidence from IMPACT. *Journal of Policy Analysis and Management 34*(2), 1–31.

Dixit, A. (2002). Incentives and Organizations in the Public Sector: An Interpretative Review. *Journal of Human Resources 37*(4), 696–727.

Dobbie, W., R. G. Fryer, and R. G. Fryer Jr (2013). Getting Beneath the Veil of Effective Schools: Evidence From New York City. *American Economic Journal: Applied Economics 5*(4), 28–60.

Duflo, E., R. Hanna, and S. P. Ryan (2012, 6). Incentives Work: Getting Teachers to Come to School. *American Economic Review 102*(4), 1241–1278.

Francois, P. (2000). 'Public Service Motivation' as an Argument for Government Provision. *Journal of Public Economics 78*(3), 275–299.

Fryer, R. G. (2013). Teacher Incentives and Student Achievement: Evidence from New York City Public Schools. *Journal of Labor Economics 31*(2), 373–407.

Gibbs, M., K. Merchant, W. Van der Steded, and M. E. Vargus (2004). Determinants and Effects of Subjectivity in Incentives. *The Accounting Review 79*(2), 409–436.

Glazerman, S. and A. Seifullah (2012). An Evaluation of the Chicago Teacher Advancement Program (Chicago TAP) after Four Years. Technical report, Mathematica Policy Research, Inc.

Glewwe, P., N. Ilias, and M. Kremer (2010, 7). Teacher Incentives. *American Economic Journal: Applied Economics 2*(3), 205–227.

Hanushek, E. A. (2007). The Single Salary Schedule and Other Issues of Teacher Pay. *Peabody Journal of Education 82*(4), 574–586.

Herman, J. L. and S. Golan (1993). The Effects of Standardized Testing on Teaching and Schools. *Educational measurement: Issues and practice 12*(4), 20–25.

Heß, S. (2017). Randomization Inference with Stata: A Guide and Software. *Stata Journal 17*(3), 630–651.

Holmstrom, B. (1982). Moral Hazard in Teams. *The Bell Journal of Economics*, 324–340.

Holmstrom, B. and P. Milgrom (1991). Multitask Principal-agent Analyses: Incentive Contracts, Asset Ownership, and Job Design. *Journal of Law, Economics & Policy 7*(24), 24–52.

Hoxby, C. M. and A. Leigh (2004). Pulled Away or Pushed Out? Explaining the Decline of Teacher Aptitude in the United States. *American Economic Review 94*(2), 236–240.

Hudson, S. (2010). The Effects of Performance-based Teacher Pay on Student Achievement. *SIEPR Discussion Papers 94305*(09), 1–49.

Imberman, S. A. and M. F. Lovenheim (2015). Incentive Strength and Teacher Productivity: Evidence from a Group-Based Teacher Incentive Pay System. *Review of Economics and Statistics 97*(2), 364–386.

Jackson, C. (2009). Student Demographics, Teacher Sorting, and Teacher Quality: Evidence from the End of School Desegregation. *Journal of Labor Economics 27*(2), 213–256.

Jacob, B. A. and S. D. Levitt (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *The Quarterly Journal of Economics 118*(3), 843–877.

Kandel, E. and E. P. Lazear (1992). Peer Pressure and Partnerships. *Journal of Political Economy 100*(4), 801–817.

Kane, T. J. and S. Cantrell (2010). Learning about Teaching: Initial Findings from the Measures of Effective Teaching Project. Technical report, Bill and Melinda Gates Foundation, MET Project, Seattle.

Kane, T. J. and D. O. Staiger (2008). Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation.

Kane, T. J. and D. O. Staiger (2012). Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains. Technical report, Bill and Melinda Gates Foundation, Met Project, Seattle.

Kleibergen, F. and R. Paap (2006). Generalized Reduced Rank Tests Using the Singular Value Decomposition. *Journal of Econometrics 133*(1), 97–126.

Koretz, D. and S. Barron (1998). The Validity of Gains in Scores on the Kentucky Instructional Results Information System. Technical report, RAND.

Kraft, M. and A. Christian (2021). Can Teacher Evaluation Systems Produce High-Quality Feedback? An Administrator Training Field Experiment (EdWorkingPaper: 19-62).

Ladd, H. F. (1999). The Dallas School Accountability and Incentive Program: An Evaluation of its Impacts on Student Outcomes. *Economics of Education Review 18*(1), 1–16.

Lankford, H., S. Loeb, and J. Wyckoff (2002, 3). Teacher Sorting and the Plight of Urban Schools: A Descriptive Analysis. *Educational Evaluation and Policy Analysis 24*(1), 37–62.

Lazear, E. P. (1986). Salaries and Piece Rates. *The Journal of Business 59*(3), 405–431.

Lazear, E. P. (2000). Performance Pay and Productivity. *American Economic Review 90*(5), 1346–1361.

Lazear, E. P. (2006, 8). Speeding, Terrorism, and Teaching to the Test. *Quarterly Journal of Economics 121*(3).

Lazear, E. P. (2012). Leadership: A Personnel Economics Approach. *Labour Economics 19*(1), 92–101.

Lazear, E. P. and S. Rosen (1981). Rank-Order Tournaments as Optimum Labor Contracts. *Journal of Political Economy 89*(5), 841–864.

Levin, J. (2003). Relational Incentive Contracts. *American Economic Review 93*(3), 835–857.

MacLeod, W. B. (2003). Optimal Contracting with Subjective Evaluation. *American Economic Review 93*(1), 216–240.

Mann, D., T. Leutscher, and R. M. Reardon (2013). Findings from a Two-year Examination of Teacher Engagement in TAP Schools across Louisianna. Technical Report September, Interactive Inc., Ashland.

Martinelli, C., S. W. Parker, A. C. Pérez-Gea, and R. Rodrigo (2018, 2). Cheating and Incentives: Learning from a Policy Experiment. *American Economic Journal: Economic Policy 10*(1), 298–325.

Mbiti, I., K. Muralidharan, M. Romero, Y. Schipper, C. Manda, and R. Rajani (2019, 8). Inputs, Incentives, and Complementarities in Education: Experimental Evidence from Tanzania*. *The Quarterly Journal of Economics 134*(3), 1627–1673.

Mbiti, I. and Y. Schipper (2021, 1). Teacher and Parental Perceptions of Performance Pay in Education: Evidence from Tanzania. *Journal of African Economies 30*(1), 55–80.

Mbiti, I. M. (2016, 8). The Need for Accountability in Education in Developing Countries. *Journal of Economic Perspectives 30*(3), 109–132.

Murnane, R. and D. Cohen (1986). Merit Pay and the Evaluation Problem: Why Most Merit Pay Plans Fail and Few Survive. *Harvard Educational Review 56*(1), 1–17.

Nagin, D. S., J. B. Rebitzer, S. Sanders, and L. J. Tayler (2002). Monitoring, Motivation, and Management: The Determinants of Opportunistic Behavior in a Field Experiment. *American Economic Review 92*(4), 850–873.

Nalbantian, H. R. and A. Schotter (1997). Productivity under Group Incentives: An Experimental Study. *American Economic Review 87*(3).

Neal, D. (2011). The Design of Performance Pay in Education. In E. Hanushek (Ed.), *Handbook of the Economics of Education* (Volume 4A ed.), Volume 4, Chapter 6, pp. 499–548. Elsevier Science.

Neill, M. (2003). The Dangers of Testing. *Educational Leadership 60*(5), 43.

Oster, E. (2019, 4). Unobservable Selection and Coefficient Stability: Theory and Evidence. *Journal of Business and Economic Statistics 37*(2), 187–204.

Pham, L. D., T. D. Nguyen, and M. G. Springer (2020, 2). Teacher Merit Pay: A Meta-Analysis. *American Educational Research Journal*.

Phipps, A. (2020). Multi-tasking with Production Uncertainty: A Real-Effort Laboratory Experiment. *Working Paper*.

Phipps, A. R. and E. A. Wiseman (2021, 4). Enacting the Rubric: Teacher Improvements in Windows of High-Stakes Observation. *Education Finance and Policy 16*(2), 283–312.

Prendergast, C. (1999). The Provision of Incentives in Firms. *Journal of Economic Literature 37*(1), 7–63.

Rivkin, S. G., E. A. Hanushek, and J. F. Kain (2005). Teachers, Schools, and Academic Achievement. *Econometrica 73*(2), 417–458.

Rockoff, J. E. (2004). The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data. *American Economic Review 94*(2), 247–252.

Sanderson, E. and F. Windmeijer (2016). A Weak Instrument F-test in Linear IV Models with Multiple Endogenous Variables. *Journal of Econometrics 190*(2), 212–221.

Sass, T. R., J. Apperson, and C. Bueno (2015). The Long Run Effects of Teacher Cheating on Student Outcomes A Report for the Atlanta Public Schools. Technical report, Georgia State University, Atlanta, GA.

Scafidi, B., D. L. Sjoquist, and T. R. Stinebrickner (2007, 4). Race, Poverty, and Teacher Mobility. *Economics of Education Review 26*(2), 145–159.

Sojourner, A. J., E. Mykerezi, and K. L. West (2014). Teacher Pay Reform and Productivity: Panel Data Evidence from Adoptions of Q-Comp in Minnesota. *Journal of Human Resources 49*(4), 945–981.

Speroni, C., A. Wellington, P. Burkander, H. Chiang, M. Herrmann, and K. Hallgren (2020, 7). Do Educator Performance Incentives Help Students? Evidence from the Teacher Incentive Fund National Evaluation. *Journal of Labor Economics 38*(3).

Springer, M. G., D. Ballou, L. Hamilton, V.-N. Le, J. R. Lockwood, D. F. Mccaffrey, M. Pepper, and B. M. Stecher (2011). Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching (POINT). Technical report.

Stecher, B. M., M. S. Garet, L. S. Hamilton, E. D. Steiner, A. Robyn, J. Poirier, D. J. Holtzman, E. S. Fulbeck, J. Chambers, and I. Brodziak De Los Reyes (2016). Improving Teaching Effectiveness. Technical report, RAND, Santa Monica, CA.

Taylor, E. S. and J. H. Tyler (2012). The Effect of Evaluation on Teacher Performance. *The American Economic Review 102*(7), 3628–3651.

Vigdor, J. L. (2008). Teacher Salary Bonuses in North Carolina. *Vanderbilt Peabody College Working Papers* (February).

Weisberg, D., S. Sexton, J. Mulhern, and D. Keeling (2009). The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness. Technical report, The New Teacher Project.

Tables

Table 1
Student Observations by Year and Grade

	2010	2011	2012	Total
4th Grade	2,246	2,035	1,958	6,239
5th Grade	1,907	2,161	1,952	6,020
Total	4,153	4,196	3,910	12,259

Sample includes all 4th and 5th grade students with a prior reading and math score whose teacher has at least one year of prior experience.

Table 2
Summary Statistics of Students by Year

	2010 (count/fraction)	2011 (count/fraction)	2012 (count/fraction)	Total (count/fraction)
Male	2,046	2,054	1,897	5,997
	0.49	0.49	0.49	0.49
White	411	487	457	1,355
	0.10	0.12	0.12	0.11
Black	2,980	2,904	2,649	8,533
	0.72	0.69	0.68	0.70
Hispanic	610	626	624	1,860
	0.15	0.15	0.16	0.15
Other Race	139	170	163	472
	0.03	0.04	0.04	0.04
English Language Learner	305	298	257	860
	0.07	0.07	0.07	0.07
Special Education	538	496	463	1,497
	0.13	0.12	0.12	0.12
Receiving Free and Reduced-Price Lunch	2,955	2,937	2,341	8,233
	0.71	0.70	0.60	0.67
N	4,153	4,196	3,910	12,259

Fraction of totals shown below counts. Sample includes all 4th and 5th grade students with a prior reading and math score whose teacher has at least one year of prior experience.

Table 3
Summary of Teacher Experience by Year

	2010 (mean/std dev)	2011 (mean/std dev)	2012 (mean/std dev)	Total (mean/std dev)
Experience	11.36	11.04	11.36	11.25
	8.17	7.95	7.88	7.99
Salary Step on Pay Scale	9.62	9.44	9.79	9.61
	4.90	4.85	4.75	4.83
N	229	243	209	681

Average Experience and Salary Pay Scale Step with standard errors below. Teachers included in the sample are all teachers with more than one year of experience in 4th and 5th grades.

Table 4
Teacher Race by Year across District

	2010 (count/fraction)	2011 (count/fraction)	2012 (count/fraction)	Total (count/fraction)
White	56	73	67	196
	0.26	0.33	0.34	0.31
Black	146	139	123	408
	0.68	0.62	0.62	0.64
Hispanic	7	6	1	14
	0.03	0.03	0.01	0.02
Asian	4	4	5	13
	0.02	0.02	0.03	0.02
Other	1	2	2	5
	0.00	0.01	0.01	0.01
N	214	224	198	636

Fraction of total included below counts. Teachers included in the sample are all teachers with more than one year of experience in 4th and 5th grades.

Table 5
Average Unmonitored Time Among Treated

	2010 (mean/std dev)	2011 (mean/std dev)	2012 (mean/std dev)	Total (mean/std dev)
Window 1	46.00	10.33	29.20	31.81
	0.00	9.58	16.36	17.81
Window 2	8.50	8.00	5.80	7.71
	6.22	5.44	6.38	5.85
Window 3	6.00	6.83	15.00	8.38
	3.23	1.72	3.46	4.73

Average unmonitored time for each window shown among teachers that experience unmonitored time. Sample is restricted to 4th and 5th grade teachers with at least one year of prior experience.

Table 6
Fraction of Teachers with Unmonitored Time by Year

	2010	2011	2012	Total
Window 1 No-Threat > 0	1.00	0.53	0.65	0.72
Window 2 Unmonitored > 0	0.26	0.16	0.18	0.20
Window 3 Unmonitored > 0	0.17	0.29	0.26	0.24
Window 4 Unmonitored > 0	0.02	0.03	0.03	0.02

Among all teachers in the sample, the fraction of those with any unmonitored time is shown with the standard deviation below. Window 4 is defined as unmonitored time *prior to the test*, which is an extremely rare event. This is why Window 4 will not be used in the analysis.

Table 7
Average Cumulative Evaluation Probability

	2010 (mean/std dev)	2011 (mean/std dev)	2012 (mean/std dev)	Total (mean/std dev)
Window 1 Cumulative Monitoring Intensity	0.00	0.81	0.75	0.52
	0.00	0.69	0.76	0.70
Window 2 Cumulative Monitoring Intensity	0.81	1.58	1.60	1.33
	0.64	0.97	1.10	0.99
Window 3 Cumulative Monitoring Intensity	1.50	1.36	1.58	1.48
	0.76	0.80	0.77	0.78

Teachers who have not yet been evaluated experience changing evaluation probability due to the changing size of the pool of remaining teachers. The daily evaluation probability can then be added across all monitored days. Evaluation probability becomes strongest in the days at the end of an evaluation window. Note that cumulative evaluation probability does not need to add up to 1. The *daily* evaluation probability must be less than 1, but not the cumulative evaluation probability.

Table 8
Average Post-Feedback Days Among Teachers Receiving Feedback Prior to Tests

	2010 (mean/std dev)	2011 (mean/std dev)	2012 (mean/std dev)	Total (mean/std dev)
P1 Post-Feedback (Days)	95.00	82.50	97.25	91.27
	1.73	4.80	11.21	9.72
P2 Post-Feedback (Days)	42.00	30.25	42.50	37.91
	7.55	4.99	18.70	12.68
P3 Post-Feedback (Days)	7.00	8.50	7.00	7.55
	2.65	6.76	3.65	4.44
M1 Post-Feedback (Days)	55.00	63.00	68.25	62.73
	8.66	27.56	22.08	20.47
M2 Post-Feedback (Days)	11.00	21.75	6.75	13.36
	2.00	13.82	2.63	10.37

Post-feedback days are calculated as the number of teaching days between when a teacher receives feedback on her in-class evaluation till the time students take their test. Notice that for $M2$ and $P3$, not all teachers receive feedback prior to standardized testing. The averages shown are calculated among teachers who received their feedback prior to testing.

Table 9
Coverage of Post-Evaluation Feedback

	2010	2011	2012	Total
Fraction with P1 Post-Feedback > 0	1.00	1.00	1.00	1.00
Fraction with P2 Post-Feedback > 0	1.00	1.00	1.00	1.00
Fraction with P3 Post-Feedback > 0	0.04	0.05	0.07	0.05
Fraction with M1 Post-Feedback > 0	1.00	1.00	1.00	1.00
Fraction with M2 Post-Feedback > 0	0.44	0.53	0.54	0.50

Because $M2$ and $P3$ may occur after student testing, not all teachers receive feedback. The fractions demonstrate how the vast majority of teachers do not receive their third Principal Evaluation until after standardized testing, while roughly half receive their second district-level evaluation before testing.

Table 10
Effect of Unmonitored Time on Student Math Test Outcomes
(Outcome: Math Standard Deviations)

	(1)	(2)	(3)	(4)	(5)
Window 1	0.0013	0.0011	0.0018	0.0014	0.0009
	(0.0013)	(0.0020)	(0.0022)	(0.0021)	(0.0021)
	[0.2042 0.2574]	[0.5628 0.6246]	[0.5276 0.5901]	[0.6175 0.6776]	[0.7541 0.8063]
Window 2	0.0018	-0.0058	-0.0054	-0.0072	-0.0092
	(0.0033)	(0.0074)	(0.0074)	(0.0072)	(0.0071)
	[0.5861 0.6472]	[0.1014 0.1428]	[0.1219 0.1662]	[0.0356 0.0631]	[0.0055 0.0196]
Window 3	-0.0105	-0.0130	-0.0134	-0.0127	-0.0112
	(0.0040)	(0.0067)	(0.0066)	(0.0063)	(0.0063)
	[0.0461 0.0766]	[0.0123 0.0307]	[0.0084 0.0246]	[0.0154 0.0355]	[0.0270 0.0518]
Previous Student Scores	X	X	X	X	X
Post-evaluation Feedback Time		X	X	X	X
Evaluation Probability Controls			X	X	X
Teacher Experience				X	X
Student Demographics					X
Observations	12305	12305	12305	12305	12305

This table demonstrates the key results for student math outcomes. All standard errors are clustered at the teacher-year level. The 95% confidence intervals for p-values from 1,000 Randomization Inference trials are shown in brackets. Sample includes students with a previous year's test score and a teacher with one or more years of experience. Coefficients for post-evaluation time and evaluation probability are included in Appendix 6. All specifications include year, school, teacher, and grade fixed-effects.

Table 11
Effect of Unmonitored Time on Student Reading Test Outcomes
(Outcome: Reading Standard Deviations)

	(1)	(2)	(3)	(4)	(5)
Window 1	0.0010	0.0037	0.0039	0.0040	0.0038
	(0.0011)	(0.0019)	(0.0019)	(0.0019)	(0.0018)
	[0.3478 0.4089]	[0.0162 0.0367]	[0.3714 0.4331]	[0.3675 0.4291]	[0.3616 0.4230]
Window 2	0.0042	0.0051	0.0049	0.0052	0.0027
	(0.0023)	(0.0053)	(0.0052)	(0.0051)	(0.0052)
	[0.0959 0.1364]	[0.0391 0.0676]	[0.0399 0.0688]	[0.0356 0.0631]	[0.2553 0.3120]
Window 3	-0.0117	-0.0144	-0.0137	-0.0142	-0.0122
	(0.0030)	(0.0052)	(0.0051)	(0.0053)	(0.0052)
	[0.0028 0.0144]	[0.0000 0.0056]	[0.0000 0.0037]	[0.0000 0.0056]	[0.0011 0.0102]
Previous Student Scores	X	X	X	X	X
Post-evaluation Feedback Time		X	X	X	X
Evaluation Probability Controls			X	X	X
Teacher Experience				X	X
Student Demographics					X
Observations	12820	12820	12820	12820	12820

This table demonstrates the key results for student reading outcomes. All standard errors are clustered at the teacher-year level. The 95% confidence intervals for p-values from 1,000 Randomization Inference trials are shown in brackets. Sample includes students with a previous year's test score and a teacher with one or more years of experience. Coefficients for post-evaluation time and evaluation probability are included in Appendix 6. All specifications include year, school, teacher, and grade fixed-effects.

Table 12
Effect of Unmonitored Time on Log Total Teacher Suspensions

	(1) Short-term Suspensions (Log)	(2) Short-term Suspensions (Log)	(3) Short-term Suspensions (Log)
Window 1	-0.000267	-0.000688	-0.000650
	(0.000379)	(0.000621)	(0.000602)
Window 2	-0.000749	0.000256	0.000370
	(0.00109)	(0.000472)	(0.000476)
Window 3	0.00251	0.00290	0.00322
	(0.00154)	(0.00157)	(0.00155)
Experience			-0.0238
			(0.0150)
Teacher FE		X	X
Observations	7071	7071	7071

Results show that students with teachers who have more unmonitored time in Window 3 are likely to have 0.3% more suspensions in the year per unmonitored day. The outcome is log student-level suspensions across the whole year. Errors are clustered at the teacher-year level. Suspension data is only available for 2011 and 2012, which results in a smaller sample size. All specifications include year, school, grade, teacher, and student fixed-effects.

Table 13
Effect of Window 3 Unmonitored on Daily Probability of Suspending a Student

	(1) Suspension Issued	(2) Suspension Issued	(3) Suspension Issued	(4) Suspension Issued
Unmonitored Day (Window 3)	0.0714	0.0732	0.0718	0.0567
	(0.0290)	(0.0291)	(0.0292)	(0.0337)
Experience		0.00510	0.00401	0.00245
		(0.00751)	(0.00781)	(0.0532)
Experience Squared		-0.000378	-0.000319	0.00238
		(0.000433)	(0.000451)	(0.00155)
School FE	X	X	X	X
Date FE	X	X	X	X
Grade FE			X	X
Teacher FE				X

This table considers the daily probability of a teacher issuing a suspension for each day in Window 3. For each teacher, there are roughly 25 days. The outcome variable for each day is a binary indicator of whether or not the teacher issued a suspension. During this time period, the average teacher issues a suspension 4.1 percent of the time. Daily suspension rates go up for unmonitored days in Window 3 by 6 to 7 percentage points.

Table 14
Placebo Test for Unmonitored Time Effect on Math Scores
(Outcome: Math Standard Deviations)

	(1)	(2)	(3)	(4)	(5)
Window 1	0.0013	0.0011	0.0018	0.0014	0.0009
	(0.0013)	(0.0020)	(0.0022)	(0.0022)	(0.0021)
Window 2	0.0016	-0.0067	-0.0062	-0.0082	-0.0101
	(0.0033)	(0.0075)	(0.0074)	(0.0072)	(0.0072)
Window 3	-0.0104	-0.0134	-0.0138	-0.0130	-0.0117
	(0.0042)	(0.0069)	(0.0069)	(0.0065)	(0.0065)
Unmonitored Placebo	-0.0010	-0.0012	-0.0011	-0.0012	-0.0011
	(0.0012)	(0.0012)	(0.0012)	(0.0012)	(0.0011)
Previous Student Scores	X	X	X	X	X
Feedback Time		X	X	X	X
Evaluation Probability Controls			X	X	X
Experience				X	X
Student Race and Gender					X
Other Student Demographics					X
Observations	12305	12305	12305	12305	12305

This test shows the results of adding a placebo treatment: unmonitored days that occur *after* students have completed their standardized tests. All errors are clustered at the teacher-year level. The key results in Windows 2 and 3 from the earlier specification remain unchanged. The placebo has no significant effect. All specifications include year, school, teacher, and grade fixed-effects.

Table 15
Placebo Test for Unmonitored Time Effect on Reading Scores
(Outcome: Reading Standard Deviations)

	(1)	(2)	(3)	(4)	(5)
Window 1	0.0010	0.0038	0.0039	0.0040	0.0038
	(0.0011)	(0.0018)	(0.0019)	(0.0019)	(0.0018)
Window 2	0.0038	0.0045	0.0046	0.0051	0.0026
	(0.0023)	(0.0052)	(0.0052)	(0.0052)	(0.0053)
Window 3	-0.0123	-0.0149	-0.0145	-0.0151	-0.0133
	(0.0030)	(0.0051)	(0.0050)	(0.0052)	(0.0051)
Unmonitored Placebo	0.0015	0.0014	0.0012	0.0012	0.0013
	(0.0010)	(0.0010)	(0.0009)	(0.0010)	(0.0009)
Previous Student Scores	X	X	X	X	X
Feedback Time		X	X	X	X
Evaluation Probability Controls			X	X	X
Experience				X	X
Student Race and Gender					X
Other Student Demographics					X
Observations	12820	12820	12820	12820	12820

This test shows the results of adding a placebo treatment: unmonitored days that occur *after* students have completed their standardized tests. All errors are clustered at the teacher-year level. The results from the earlier specification remain unchanged, and the placebo has no significant effect. All specifications include year, school, teacher, and grade fixed-effects.

Table 16
Proportional Selection on Unobservables Relative to Observables Required to Obtain Observed Effect Sizes under Different Maximum R-Squared Values

R-Squared	Math	Reading
0.70		12.16
0.75	119.81	3.44
0.80	58.20	2.00
0.85	38.43	1.41
0.90	28.69	1.09
0.95	22.89	0.89

These are the results the robustness tests suggested in Altonji et al. (2005) for the effects of unmonitored time in Window 3. For an assumed maximum R^2 value (Column 1), results show the required amount of selection on unobservables to achieve the observed coefficients as a fraction of how much observables explain the outcome. The recommendation from Altonji et al. (2005) is that if 100% or more selection on unobservables is required then the results are considered robust. Values are calculated using `psaCalc` in Oster (2019).

Table 17
Changes in Evaluation Score under Less Monitoring
(Outcome: Evaluation Points)

		Decreased Monitoring	
		M2 Pre-test	M2 Post-test
Standard 1	Lead well-organized, objective-driven lessons	0.390 (1.039)	-1.116 (1.542)
Standard 2	Explain content clearly	0.058 (0.658)	0.836 (0.903)
Standard 3	Engage students at all learning levels in accessible and challenging work	-0.921 (0.717)	-0.837 (1.008)
Standard 4	Provide students multiple ways to move toward mastery	-0.707 (0.942)	0.158 (0.996)
Standard 5	Check for student understanding	0.746 (0.454)	-0.632 (0.998)
Standard 6	Scaffolding, probing, and re-teaching	-0.863 (0.441)	0.864 (0.645)
Standard 7	Develop higher-level understanding through effective questioning	-1.495 (0.672)	-0.610 (0.967)
Standard 8	Maximize instructional time (including pacing, student behavior and idleness)	-1.305 (0.562)	0.745 (0.880)
Standard 9	Build a supportive, learning-focused classroom community	0.123 (0.476)	-0.275 (0.850)
N		413	259

This table measures how evaluation scores change as the probability of an evaluation decreases for evaluations occurring before and after standardized testing. Confidence levels are not adjusted for multiple hypothesis testing. The specification for each regression is $S_{ijs}^{M2} = \boldsymbol{X}_{ij}\Gamma - p_{ij}\mu + \bar{p}_{ij}\bar{\mu} + \sum_{q=P1,M1,P2} S_{ij}^{q}\nu^{q} + \boldsymbol{T}_{ij}\omega + \phi_s + \delta_j + \varepsilon_{ijs}$, and estimates for μ are shown in the table (see Equation 3). $\boldsymbol{X}_{ij}$ is a vector of experience indicators, ϕ_s is a school fixed-effect and δ_j is a year fixed-effect. $\boldsymbol{T}_{ij}$ is a vector of indicators for if the M2 evaluation was third (before P2), fourth (after P2), or fifth (after P3). S_{ij}^{q} are scores on standard S for $q = P1, M1, P2$, which are evaluations that must occur before testing. The term $p_{ij} = Pr(\text{Eval})$ is measured as the probability of receiving an evaluation on the day of the M2 evaluation. The term $\bar{p}_{ij} = \sum_{d=1}^{D-1} Pr(\text{Eval})_{ijd}$ is the sum of the daily probability of an M2 evaluation. The results demonstrate how in the time leading up to standardized tests, teachers sacrifice pacing lessons appropriately, managing student behavior, and maintaining focus (Items 7 and 8) when they are monitored less.

Table 18
Effect of Unmonitored Time on
Student Math Score Dispersion

	(1) Δ 5-95	(2) Δ 10-90	(3) Δ 20-80	(4) Δ 30-70
Window 1	-0.00145	0.00114	-0.00028	-0.00018
	(0.00251)	(0.00169)	(0.00123)	(0.00093)
	[0.484 0.546]	[0.441 0.503]	[0.784 0.834]	[0.814 0.860]
Window 2	0.01320	0.01191	0.00534	0.00191
	(0.00823)	(0.00543)	(0.00437)	(0.00303)
	[0.034 0.061]	[0.008 0.025]	[0.070 0.106]	[0.416 0.478]
Window 3	-0.01620	-0.00627	-0.00230	-0.00297
	(0.00798)	(0.00516)	(0.00381)	(0.00233)
	[0.016 0.037]	[0.215 0.269]	[0.500 0.562]	[0.250 0.307]
Observations	452	452	452	452

Coefficients show how much inter-percentile distance changes for student math scores as a function of unmonitored days for each window. Distances are 5th to 95th percentile, 10th to 90th percentile, 20th to 80th percentile, and 30th to 70th percentile. The 95% confidence interval for p-values from 1,000 randomization inference trials are shown in brackets and the sample is the same as in Table 10. Several of the coefficients in Window 3 are statistically significant, and all the coefficients are consistently negative in Window 3 but not in other windows. This coincides with teachers reducing the distribution of their student test scores during unmonitored time as standardized tests approach.

Table 19
Effect of Unmonitored Time on
Student Reading Score Dispersion

	(1) Δ 5-95	(2) Δ 10-90	(3) Δ 20-80	(4) Δ 30-70
Window 1	0.00060	0.00239	0.00105	-0.00005
	(0.00258)	(0.00170)	(0.00119)	(0.00088)
	[0.773 0.823]	[0.115 0.159]	[0.368 0.430]	[0.928 0.957]
Window 2	0.02190	0.00701	0.00477	0.00024
	(0.00884)	(0.00600)	(0.00484)	(0.00258)
	[0.000 0.007]	[0.102 0.144]	[0.134 0.180]	[0.897 0.932]
Window 3	-0.01915	-0.00781	-0.00773	-0.00349
	(0.00880)	(0.00514)	(0.00385)	(0.00243)
	[0.005 0.018]	[0.113 0.157]	[0.035 0.062]	[0.201 0.254]
Observations	452	452	452	452

Coefficients show how much inter-percentile distance changes for student math scores as a function of unmonitored days for each window. Distances are 5th to 95th percentile, 10th to 90th percentile, 20th to 80th percentile, and 30th to 70th percentile. The 95% confidence interval for p-values from 1,000 randomization inference trials are shown in brackets and the sample is the same as in Table 11. Several of the coefficients in Window 3 are statistically significant, and all the coefficients are consistently negative in Window 3 but not in other windows. This coincides with teachers reducing the distribution of their student test scores during unmonitored time as standardized tests approach.

Table 20
Effect of Unmonitored Time on Student Math Outcomes
Without Value-Added Incentive (3rd Grade)
(Outcome: Math Standard Deviations)

	(1)	(2)	(3)	(4)	(5)
Window 1	0.0001	-0.0014	-0.0027	-0.0021	-0.0022
	(0.0031)	(0.0039)	(0.0035)	(0.0032)	(0.0026)
	[0.868 0.908]	[0.843 0.887]	[0.812 0.858]	[0.827 0.873]	[0.746 0.799]
Window 2	0.0065	0.0179	0.0278	0.0232	0.0157
	(0.0049)	(0.0135)	(0.0141)	(0.0135)	(0.0139)
	[0.376 0.438]	[0.041 0.070]	[0.000 0.007]	[0.008 0.023]	[0.746 0.799]
Window 3	0.0192	0.0062	-0.0009	-0.0059	-0.0021
	(0.0051)	(0.0105)	(0.0097)	(0.0090)	(0.0067)
	[0.039 0.068]	[0.534 0.596]	[0.907 0.941]	[0.549 0.611]	[0.695 0.752]
School FE	X	X	X	X	X
Post-evaluation Feedback Time		X	X	X	X
Evaluation Probability Controls			X	X	X
Experience				X	X
Student Race and Gender					X
Other Student Demographics					X
Observations	2529	2529	2529	2529	2529

This table demonstrates the key results for 3rd grade student math outcomes during the 2011-2012 school year. 3rd grade teachers do not receive test-based bonuses. All standard errors are clustered at the teacher-year level. The sample includes students with a teacher with one or more years of experience. Because only one year of data is available, there are no teacher fixed effects and no previous student test scores.

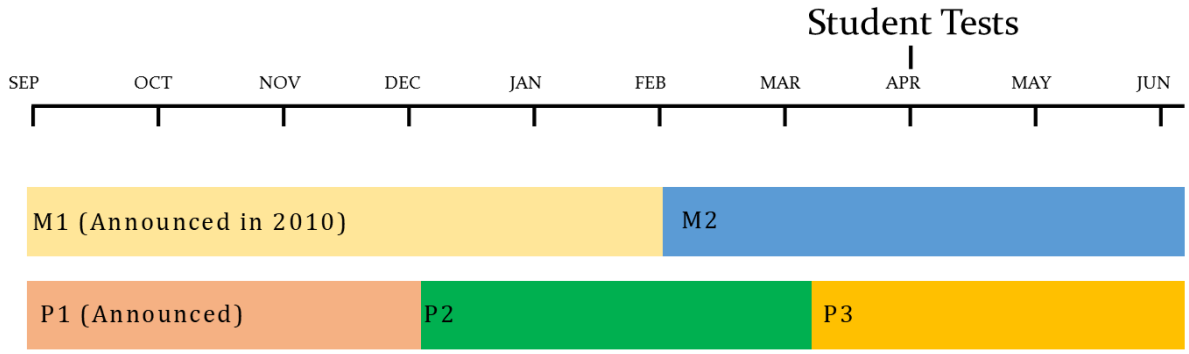
Table 21
Effect of Unmonitored Time on Student Reading Outcomes
Without Value-Added Incentive (3rd Grade)
(Outcome: Reading Standard Deviations)

	(1)	(2)	(3)	(4)	(5)
Window 1	0.0018	0.0003	0.0033	0.0050	0.0010
	(0.0023)	(0.0033)	(0.0040)	(0.0048)	(0.0038)
	[0.565 0.627]	[0.962 0.983]	[0.603 0.664]	[0.491 0.553]	[0.782 0.832]
Window 2	0.0027	0.0104	0.0284	0.0279	0.0155
	(0.0034)	(0.0173)	(0.0196)	(0.0192)	(0.0168)
	[0.692 0.749]	[0.198 0.250]	[0.000 0.006]	[0.002 0.012]	[0.782 0.832]
Window 3	-0.0094	-0.0063	-0.0114	-0.0135	-0.0134
	(0.0130)	(0.0148)	(0.0149)	(0.0154)	(0.0121)
	[0.278 0.336]	[0.471 0.533]	[0.236 0.291]	[0.190 0.242]	[0.305 0.364]
School FE	X	X	X	X	X
Post-evaluation Feedback Time		X	X	X	X
Evaluation Probability Controls			X	X	X
Experience				X	X
Student Race and Gender					X
Other Student Demographics					X
Observations	2443	2443	2443	2443	2443

This table demonstrates the key results for 3rd grade student reading outcomes during the 2011-2012 school year. 3rd grade teachers do not receive test-based bonuses. All standard errors are clustered at the teacher-year level. The sample includes students with a teacher with one or more years of experience. Because only one year of data is available, there are no teacher fixed effects and no previous student test scores.

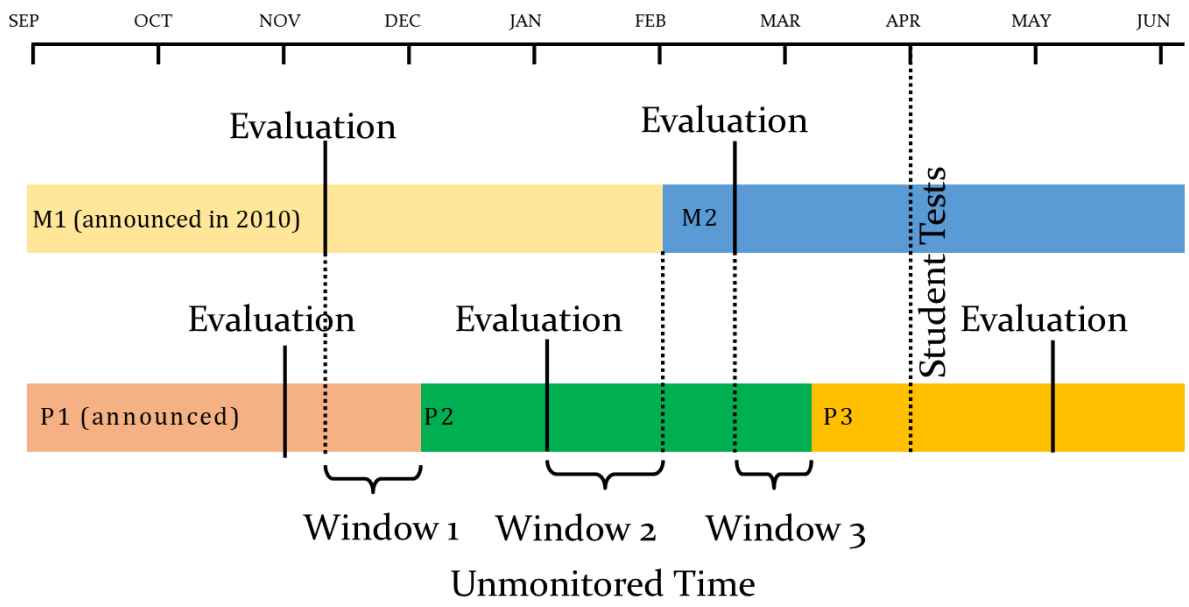
Figures

Figure 1
Depiction of each evaluation window in DCPS



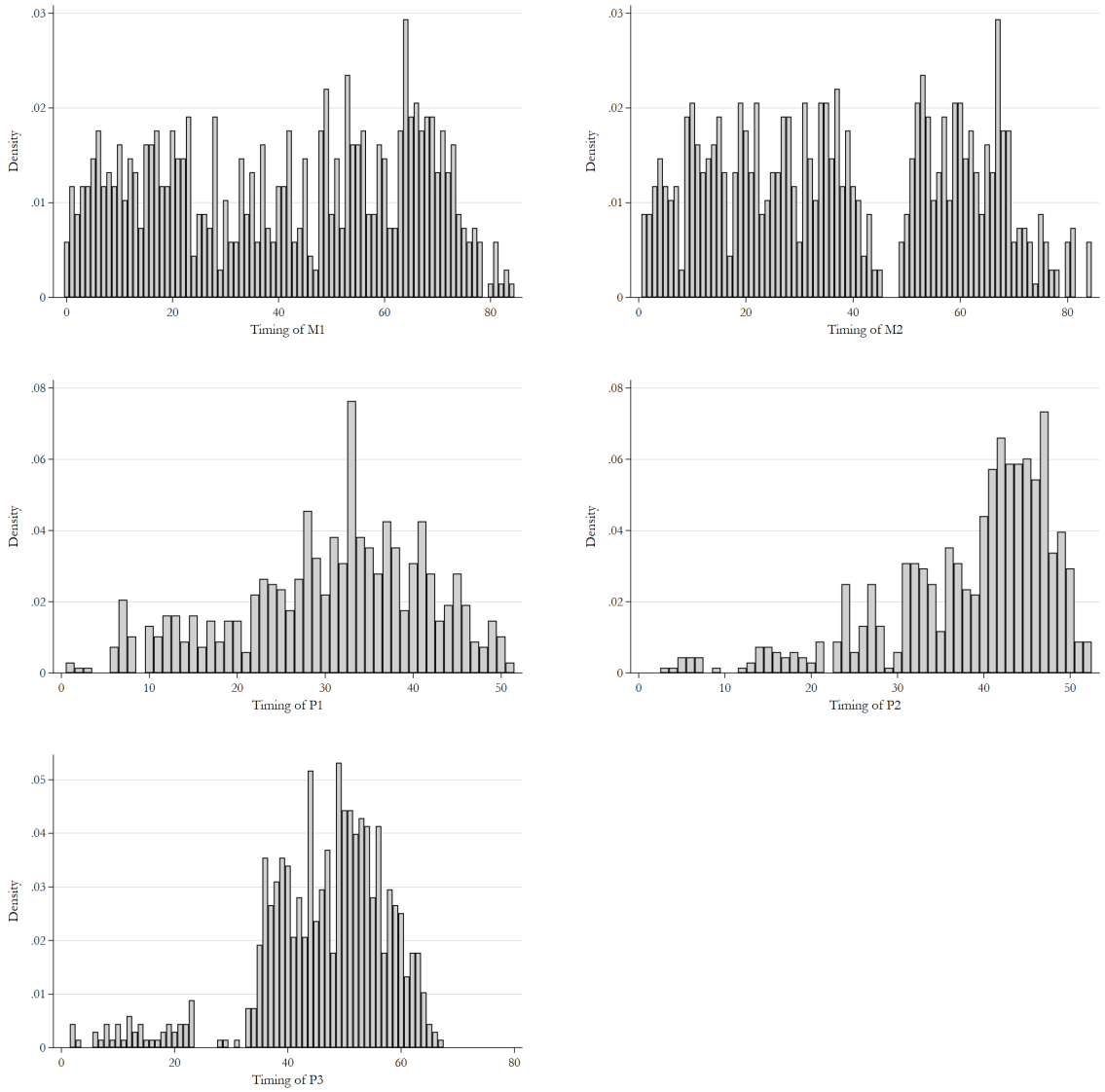
Note: One evaluation must occur within each window. For announced evaluations, teachers are informed no later than the day before their evaluation

Figure 2
Calculating unmonitored time in DCPS



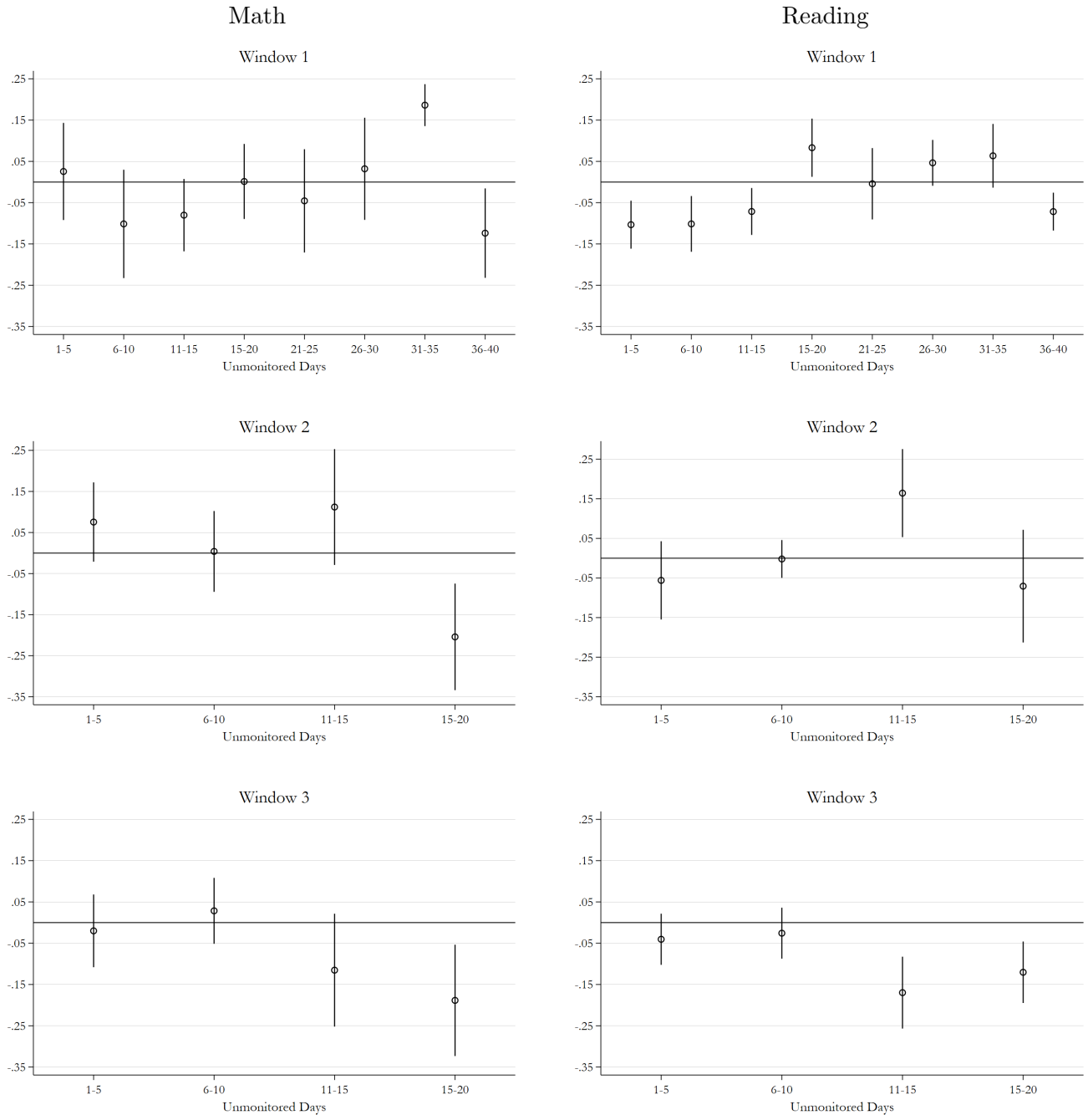
Note: Because evaluation windows in DCPS overlap, unmonitored time is defined as days in which there is no possibility of an unannounced evaluation from either evaluator. Window 1 is possible between September and December, when the second principal evaluation window starts. Window 2 starts December 1 and ends February 1, and Window 3 starts February 1 and ends March 15. Unmonitored time is not calculated for $P3$ because very few teachers have any unmonitored time prior to the test (19 instances over all years, or less than 4%).

Figure 3
Timing of Evaluations within Evaluation Window



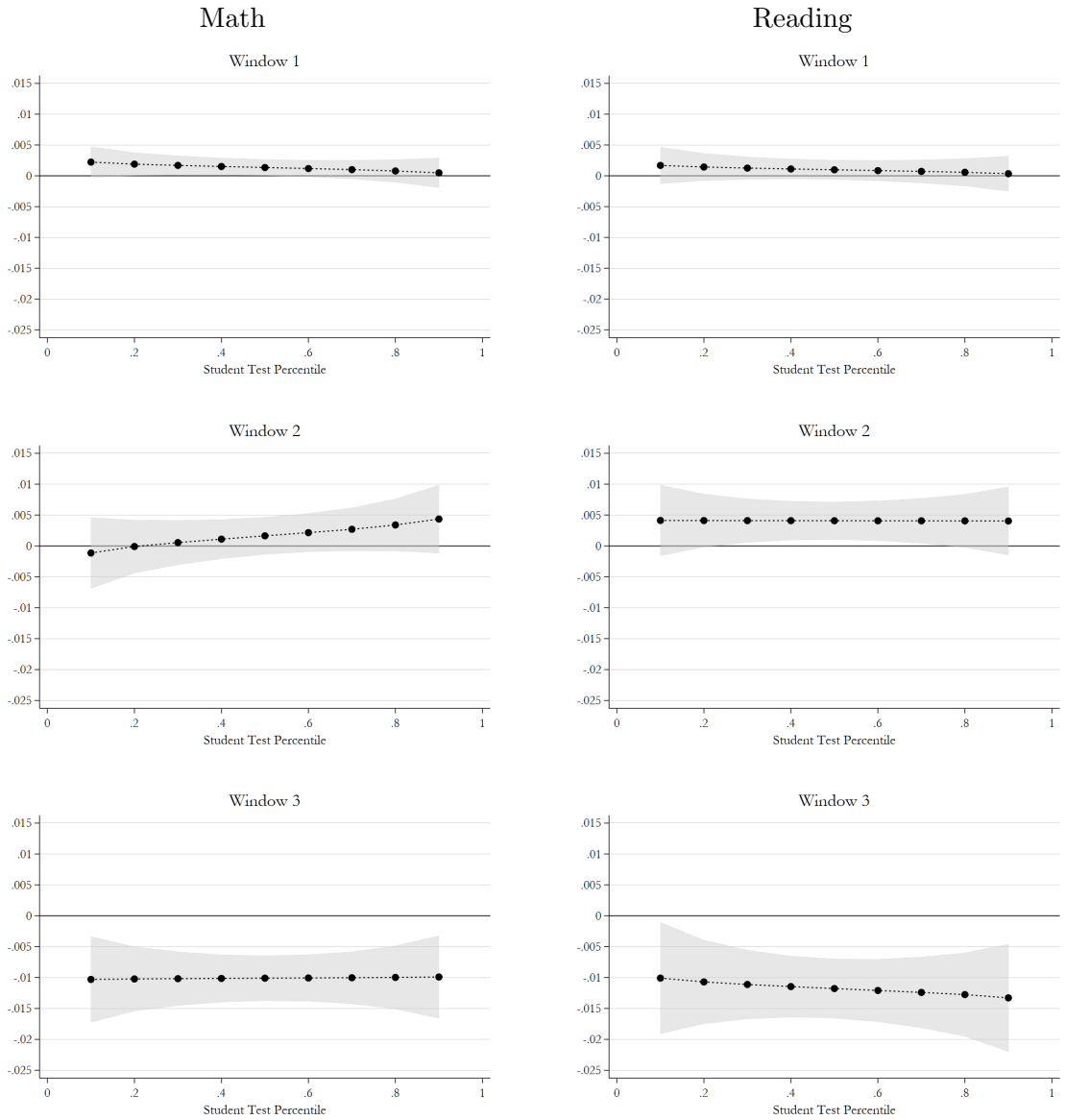
Note: Days are measured as instruction days, which excludes in-service days, weekends, and holidays. Master educator evaluations, $M1$ and $M2$, are distributed uniformly across the window. Principal evaluations – $P1$, $P2$ and $P3$ – are often clustered near the end of each window.

Figure 4
Non-Linear Cumulative Effect Sizes in Math and Reading



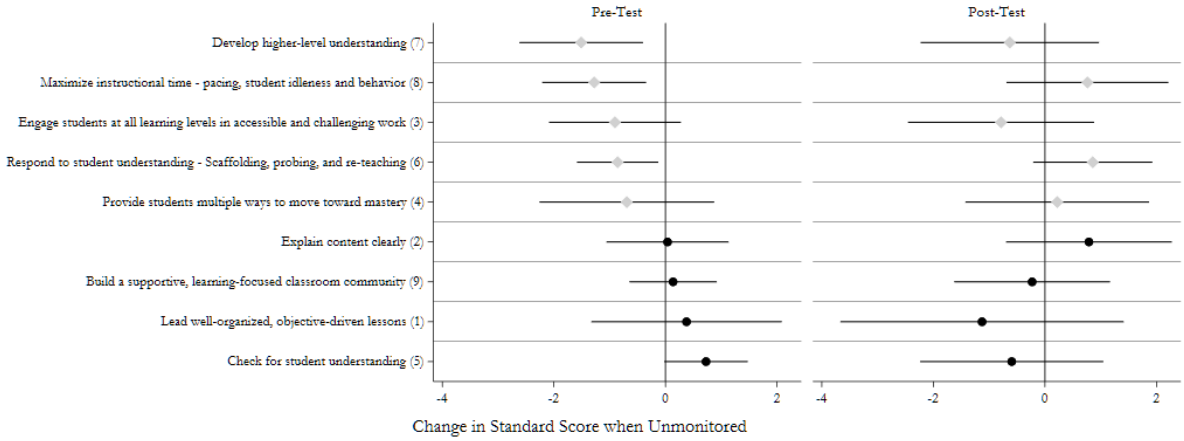
Note: The figures depict the estimated non-linear effects of unmonitored days in each window. For Window 3, the majority of effects appear during the third week of unmonitored time. Bands indicate 90% confidence intervals.

Figure 5
Effects by Student Decile



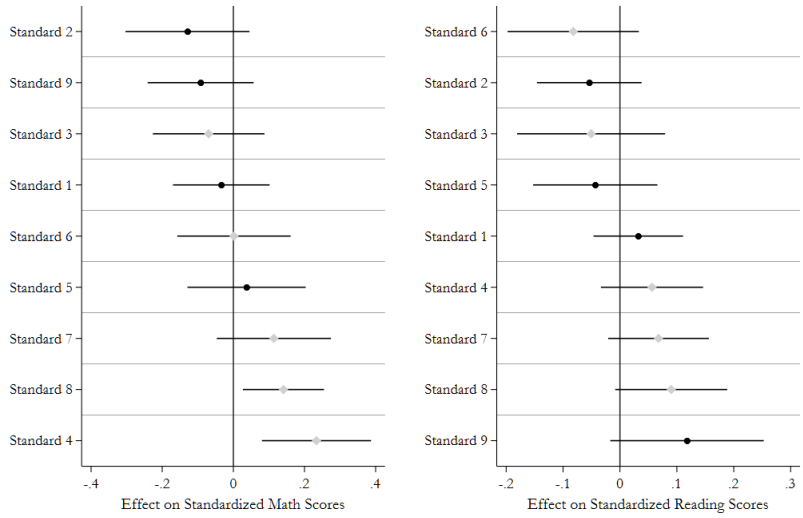
Note: Results are from quantile regression based on student test decile from the previous year. The results here are consistent with the overall results: Window 3 has statistically significant and negative effect. However, there do not appear to be any meaningful differences between students based on previous test performance.

Figure 6
Effect of Decreasing Monitoring on Evaluation Score



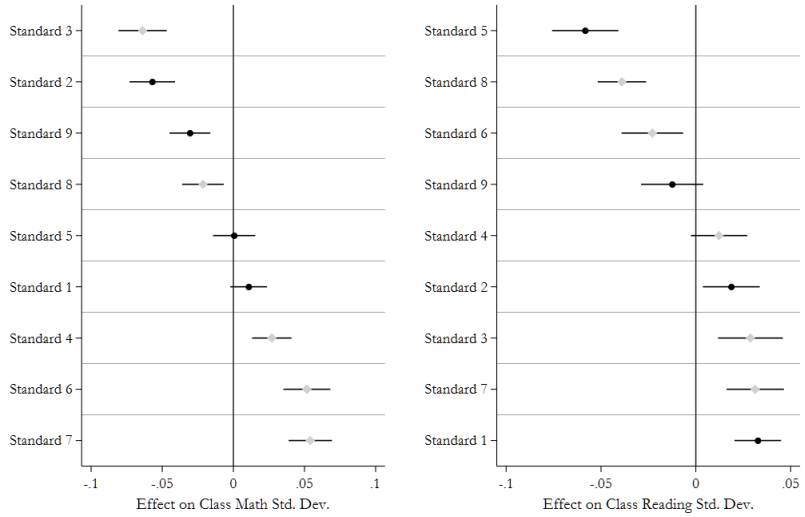
Note: This figure reports the estimated effect of reducing monitoring. The outcome is a teacher's score for each Standard (results are shown in Table 17). Lines indicate 90% confidence intervals. See Table A27 for a full description of the Standards. Standards 3, 4, 6, 7, 8 are the Standards we would expect to suffer the most if teachers are teaching to the test (shown with gray dots). Before a test, teachers seem to sacrifice these standards when they are monitored less, but only before standardized testing. After tests, there are no clear patterns of change when monitored less. The specification is $S_{ijs}^{M2} = \boldsymbol{X}_{ij}\Gamma - p_{ij}\mu + \bar{p}_{ij}\bar{\mu} + \sum_q S_{ij}^q \nu^q + \boldsymbol{T}_{ij}\omega + \phi_s + \delta_j + \varepsilon_{ijs}$ where values of μ are illustrated. $\boldsymbol{X}_{ij}$ is a vector of experience indicators, ϕ_s is a school fixed-effect and δ_j is a year fixed-effect (see Equation 3). $\boldsymbol{T}_{ij}$ is a vector of indicators for if the $M2$ evaluation was third (before $P2$), fourth (after $P2$), or fifth (after $P3$). S_{ij}^q are scores on standard S for $q = P1, M1, P2$, which are evaluations that must occur before testing. The term $p_{ij} = Pr(\text{Eval})$ is measured as the probability of receiving an evaluation on the day of the $M2$ evaluation. The term $\bar{p}_{ij} = \sum_{d=1}^{D-1} Pr(\text{Eval})_{ijd}$ is the sum of the daily probability of an $M2$ evaluation and D indicates the day of the evaluation.

Figure 7
Effect of Teaching Standards on Student Scores (IV)



Note: This figure shows the estimated effect of improving each teaching Standard on student test scores. The specification for student k with teacher i in year j is $Y_{kijs} = W_{kj}\Omega + X_{ij}\Gamma + \hat{\boldsymbol{S}}_{ijs}^{M2}\eta + \sum_{w=1}^{3}\beta^{w}n_{ij}^{w} + \phi_s + \delta_i + \varepsilon_{kijs}$, where $Z = \left[\boldsymbol{S}_{ijs}^{\mathrm{P1}}, \boldsymbol{S}_{ijs}^{\mathrm{M1}}, \boldsymbol{S}_{ijs}^{\mathrm{P2}}, \boldsymbol{T}_{ij}, p_{ij}, \bar{p}_{ij}\right]$ are instruments for $\hat{\boldsymbol{S}}_{ijs}^{M2}$ (see Equation 4). The terms n_{ij}^{w} are, as before, the number of unmonitored days for each window w. The student variables in W_{kj} include gender, race, ELL status, free and reduced-price lunch status, and previous student test scores. Lines indicate 90% confidence intervals. Gray dots indicate standards expected to suffer when a teacher is teaching to the test. Based on these estimates, the Standards teachers sacrifice most in the time before standardized testing are among the most effective. The estimates should be interpreted with caution, however. Using the joint test for under-identification using the rK statistic from Kleibergen and Paap (2006), the model fails to reject the null-hypothesis that the endogenous variables are under-identified. Using the weak identification test (Kleibergen-Paap Wald rK F statistic), the endogenous variables are weakly identified. This is somewhat expected given the well-established collinearity between evaluation standards (see Adnot et al., 2017). However, when identification is tested separately for each endogenous variable using the method in Sanderson and Windmeijer (2016), the model is well-identified.

Figure 8
Effect of Teaching Standards on Score Dispersion (IV)



Note: This figure displays the effect of each evaluation Standard (Standards 1-9) on the standard deviation of a teacher's student test scores using an instrumental variables approach. Evaluation scores are from the M2 evaluation and are measured in standard deviations by component. The specification is $D_{ijs} = X_{ij}\Gamma + D_{ijs}^{\text{prev}}\rho + \hat{\boldsymbol{S}}_{iys}^{M2}\eta + \sum_{w=1}^{3}\beta^{w}n_{ij}^{w} + \delta_j + \varepsilon_{ijs}$ where $Z = \left[\boldsymbol{S}_{ijs}^{\text{P1}}, \boldsymbol{S}_{ijs}^{\text{M1}}, \boldsymbol{S}_{ijs}^{\text{P2}}, \boldsymbol{T}_{ij}, p_{ij}, \bar{p}_{ij}\right]$ are instruments for $\hat{\boldsymbol{S}}_{ijs}^{M2}$ (see Equation 5). The term D_{ijs}^{prev} is the standard deviation for a teacher's current students but calculated using their previous scores from the year before. The terms n_{ij}^{w} are, as before, the number of unmonitored days for each window w. Lines indicate 90% confidence intervals. Gray dots indicate standards expected to suffer when a teacher is teaching to the test. Standard 7 has the highest effect on student score dispersion for both reading and math, which is predictable given Standard 7 measures a teacher's effort to teach "higher-level learning." Standards 8 and 3 concern teaching to all students at their level and classroom management, which are unlikely to increase student score dispersion. Standard 4 measures a teacher's use of multiple learning methods, while Standard 6 measures how well a teacher builds underlying understanding, allowing students to build upon their own learning (scaffolding). Using the joint test for under-identification using the rK statistic from Kleibergen and Paap (2006), the results reject the null-hypothesis that the endogenous variables are under-identified. Using the weak identification test (Kleibergen-Paap Wald rK F statistic), the endogenous variables are not weakly identified. Similarly, when identification is tested separately for each endogenous variable using the method in Sanderson and Windmeijer (2016), the model is well-identified.

Online Appendix

Additional Summary Statistics

Table A1
Average School-level Student Demographics

	2010 (mean/std dev)	2011 (mean/std dev)	2012 (mean/std dev)	Total (mean/std dev)
Fraction White	0.06	0.07	0.07	0.07
	0.16	0.18	0.18	0.17
Fraction Black	0.79	0.77	0.77	0.77
	0.30	0.31	0.30	0.30
Fraction Hispanic	0.12	0.13	0.13	0.13
	0.21	0.21	0.20	0.21
Fraction Other Race	0.02	0.03	0.03	0.03
	0.05	0.05	0.05	0.05
Fraction English Learner	0.06	0.06	0.05	0.06
	0.13	0.12	0.10	0.12
Fraction Special Education	0.13	0.13	0.12	0.13
	0.07	0.12	0.08	0.09
Fraction Free or Reduced-Price Lunch	0.75	0.77	0.66	0.73
	0.29	0.28	0.24	0.27
Number of Students	51.27	54.49	48.27	51.29
	27.30	30.10	25.79	27.74
N	81	77	81	239

School-level averages of student composition with standard deviation below. Sample includes all 4th and 5th grade students with a prior reading and math score whose teacher has at least one year of prior experience.

Table A2
Average School-level Teacher Demographics

	2010 (mean/std dev)	2011 (mean/std dev)	2012 (mean/std dev)	Total (mean/std dev)
Average Teacher Experience	11.13	11.11	11.65	11.30
	6.11	5.53	6.11	5.91
Average Career Step	9.37	9.46	9.90	9.57
	3.80	3.43	3.65	3.62
Fraction White	0.21	0.24	0.29	0.25
	0.33	0.31	0.36	0.33
Fraction Black	0.67	0.63	0.62	0.64
	0.35	0.35	0.38	0.36
Fraction Hispanic	0.03	0.01	0.00	0.01
	0.10	0.06	0.02	0.07
Asian	0.01	0.02	0.02	0.02
	0.07	0.09	0.06	0.07
Average Number of Teachers	2.86	3.20	2.71	2.92
	1.49	1.80	1.21	1.52
N	80	76	77	233

School-level average with standard deviation below. Teachers included in the sample are all teachers with more than one year of experience in 4th and 5th grades.

Balance Checks

Table A3
Balance Check between No-Treatment and Any-Treatment across Students

	No Treatment (Window 1)	Any Treatment (Window 1)	Diff	Std Error	P-Value	Obs
Male	0.0001	−0.0001	0.0002	0.0079	0.9845	16 119
Black	0.0023	−0.0031	0.0054	0.0051	0.2890	16 119
Hispanic	−0.0026	0.0036	−0.0061	0.0047	0.1919	16 119
Other Race	−0.0001	0.0002	−0.0003	0.0032	0.9227	16 119
English Learner	−0.0031	0.0043	−0.0075*	0.0039	0.0552	16 119
Special Education	−0.0024	0.0032	−0.0056	0.0050	0.2621	16 119
Free or Reduced-Price Lunch	0.0000	0.0000	0.0000	0.0058	0.9992	16 119
Previous Math Score	0.0016	−0.0022	0.0038	0.0126	0.7638	16 119
Previous Reading Score	0.0008	−0.0011	0.0020	0.0128	0.8778	16 119

	No Treatment (Window 2)	Any Treatment (Window 2)	Diff	Std Error	P-Value	Obs
Male	−0.0004	0.0017	−0.0021	0.0099	0.8342	16 119
Black	−0.0018	0.0075	−0.0092	0.0064	0.1493	16 119
Hispanic	0.0016	−0.0068	0.0084	0.0058	0.1460	16 119
Other Race	−0.0001	0.0003	−0.0003	0.0038	0.9292	16 119
English Learner	0.0005	−0.0020	0.0025	0.0049	0.6071	16 119
Special Education	−0.0015	0.0064	−0.0080	0.0066	0.2265	16 119
Free or Reduced-Price Lunch	0.0001	−0.0003	0.0003	0.0074	0.9647	16 119
Previous Math Score	−0.0016	0.0068	−0.0084	0.0163	0.6078	16 119
Previous Reading Score	−0.0004	0.0015	−0.0018	0.0164	0.9106	16 119

	No Treatment (Window 3)	Any Treatment (Window 3)	Diff	Std Error	P-Value	Obs
Male	−0.0003	0.0010	−0.0013	0.0090	0.8876	16 119
Black	0.0008	−0.0025	0.0033	0.0053	0.5317	16 119
Hispanic	−0.0003	0.0010	−0.0013	0.0050	0.7899	16 119
Other Race	0.0000	0.0000	0.0001	0.0032	0.9854	16 119
English Learner	−0.0009	0.0028	−0.0037	0.0044	0.4081	16 119
Special Education	0.0011	−0.0033	0.0044	0.0057	0.4424	16 119
Free or Reduced-Price Lunch	−0.0007	0.0023	−0.0030	0.0065	0.6450	16 119
Previous Math Score	−0.0002	0.0006	−0.0008	0.0143	0.9528	16 119
Previous Reading Score	0.0011	−0.0033	0.0044	0.0147	0.7641	16 119

This table compares the average student characteristics across those receiving any treatment (teacher has an unmonitored day) versus those who experience no treatment (teacher is always monitored in the given Window). Values shown in columns 1 and 2 are deviations from the school mean. In other words, these are average within-school deviations between treated and untreated. The only marginally significant differences is in Window 1, where it appears that English Language Learners are marginally more likely to have some treatment. Because treatment occurs at the teacher level, all student-teacher combinations are preserved which is why observations are greater than in Table 1.

Table A4
Balance Check for Window 1 across Students
(Outcome: Window 1 Unmonitored Days)

	(1)	(2)	(3)	(4)
Teacher Experience	-0.0921	-0.0925	-0.0914	-0.0925
	(0.180)	(0.180)	(0.180)	(0.180)
Male		0.0655	0.0770	0.115
		(0.113)	(0.112)	(0.115)
Black		-0.246	-0.275	-0.320
		(0.374)	(0.397)	(0.388)
Hispanic		-0.364	-0.479	-0.499
		(0.476)	(0.456)	(0.454)
Other Race		-0.0791	-0.140	-0.130
		(0.417)	(0.422)	(0.422)
English Learner			0.282	0.289
			(0.332)	(0.324)
Special Education			-0.271	-0.273
			(0.428)	(0.418)
Free or Reduced Price Lunch			0.229	0.230
			(0.306)	(0.308)
Previous Reading				0.224*
				(0.136)
Previous Math				-0.264*
				(0.152)
School FE	X	X	X	X
Year FE	X	X	X	X
Subject FE	X	X	X	X
Observations	16119	16119	16119	16119
F	0.463	0.397	0.503	0.813
p	0.630	0.881	0.873	0.627

This table checks if treatment is systematically targeted at certain students based on observable characteristics. Errors are clustered at the teacher-year level. The outcome is number of unmonitored days in Window 1 for a student's teacher. There do not appear to be any statistically significant coefficients. The F-Test is a joint hypothesis test of whether all coefficients are zero (not including the fixed-effects). These too appear to fail to reject the null hypothesis. Because treatment occurs at the teacher level, all student-teacher combinations are preserved which is why observations are greater than in Table 1.

Table A5
Balance Check for Window 2 across Students
(Outcome: Window 2 Unmonitored Days)

	(1)	(2)	(3)	(4)
Teacher Experience	0.0955	0.0953	0.0942	0.0946
	(0.0836)	(0.0837)	(0.0836)	(0.0836)
Male		0.0119	0.0304	0.0224
		(0.0588)	(0.0533)	(0.0556)
Black		0.000241	0.0500	0.108
		(0.129)	(0.131)	(0.133)
Hispanic		-0.157	-0.0681	-0.0345
		(0.144)	(0.142)	(0.146)
Other Race		0.0154	0.0493	0.0514
		(0.176)	(0.176)	(0.176)
English Learner			-0.141	-0.0894
			(0.173)	(0.172)
Special Education			-0.280	-0.218
			(0.376)	(0.354)
Free or Reduced Price Lunch			-0.196	-0.184
			(0.125)	(0.127)
Previous Reading				-0.0382
				(0.0662)
Previous Math				0.133*
				(0.0701)
School FE	X	X	X	X
Year FE	X	X	X	X
Subject FE	X	X	X	X
Observations	16119	16119	16119	16119
F	2.604	1.454	1.088	1.317
p	0.0747	0.192	0.369	0.210

This table checks if treatment is systematically targeted at certain students based on observable characteristics. Errors are clustered at the teacher-year level. The outcome is number of unmonitored days in Window 2 for a student's teacher. There do not appear to be any statistically significant coefficients except the student's previous math score is marginally significant (at the 10% level). Because the coefficient is positive, this would positively bias my estimate of the effect of unmonitored days. The F-Test is joint hypothesis test of whether all coefficients are zero (not including the fixed-effects). These too appear to fail to reject the null hypothesis. Because treatment occurs at the teacher level, all student-teacher combinations are preserved which is why observations are greater than in Table 1.

Table A6
Balance Check Window 3 across Students
(Outcome: Window 3 Unmonitored Days)

	(1)	(2)	(3)	(4)
Teacher Experience	-0.00228	-0.00217	-0.00135	-0.00140
	(0.0691)	(0.0690)	(0.0689)	(0.0688)
Male		0.0432	0.0341	0.0416
		(0.0398)	(0.0398)	(0.0396)
Black		0.0230	-0.0147	0.0260
		(0.0817)	(0.0849)	(0.0846)
Hispanic		0.222	0.158	0.184
		(0.154)	(0.155)	(0.158)
Other Race		0.0602	0.0332	0.0395
		(0.113)	(0.116)	(0.116)
English Learner			0.0877	0.144
			(0.0845)	(0.0892)
Special Education			0.131	0.192
			(0.144)	(0.146)
Free or Reduced Price Lunch			0.154	0.167
			(0.102)	(0.103)
Previous Reading				0.0533
				(0.0467)
Previous Math				0.0263
				(0.0528)
School FE	X	X	X	X
Year FE	X	X	X	X
Subject FE	X	X	X	X
Observations	16119	16119	16119	16119
F	0.215	0.646	0.826	0.973
p	0.807	0.693	0.592	0.469

This table checks if treatment is systematically targeted at certain students based on observable characteristics. Errors are clustered at the teacher-year level. The outcome is number of unmonitored days in Window 3 for a student's teacher. There do not appear to be any statistically significant coefficients. The F-Test is a joint hypothesis test of whether all coefficients are zero (not including the fixed-effects). These too appear to fail to reject the null hypothesis. Because treatment occurs at the teacher level, all student-teacher combinations are preserved which is why observations are greater than in Table 1.

Table A7
Balance between No-Treatment and Any-Treatment across Teachers

	No Treatment (Window 1)	Any Treatment (Window 1)	Diff	Std Error	P-Value	Obs
Prior Experience	−0.1535	0.2517	−0.4052	0.4766	0.3956	681
Black	−0.0036	0.0059	−0.0096	0.0272	0.7254	681
Hispanic	−0.0047	0.0078	−0.0125	0.0094	0.1841	681
White	0.0070	−0.0114	0.0184	0.0247	0.4559	681
Salary Step	−0.0499	0.0818	−0.1317	0.2803	0.6387	681
Evaluatoin Score (t-1)	−0.0186	0.0146	−0.0332	0.0278	0.2320	405

	No Treatment (Window 2)	Any Treatment (Window 2)	Diff	Std Error	P-Value	Obs
Prior Experience	−0.0054	0.0216	−0.0270	0.6107	0.9648	681
Black	0.0045	−0.0179	0.0224	0.0352	0.5265	681
Hispanic	0.0005	−0.0018	0.0023	0.0120	0.8484	681
White	−0.0063	0.0254	−0.0317	0.0288	0.2725	681
Salary Step	0.0279	−0.1117	0.1396	0.3564	0.6958	681
Evaluatoin Score (t-1)	0.0090	−0.0416	0.0506	0.0357	0.1598	405

	No Treatment (Window 3)	Any Treatment (Window 3)	Diff	Std Error	P-Value	Obs
Prior Experience	0.0295	−0.0930	0.1225	0.5482	0.8234	681
Black	−0.0044	0.0140	−0.0185	0.0303	0.5421	681
Hispanic	0.0007	−0.0022	0.0029	0.0098	0.7645	681
White	0.0051	−0.0160	0.0210	0.0273	0.4422	681
Salary Step	0.0196	−0.0617	0.0813	0.3311	0.8063	681
Evaluatoin Score (t-1)	−0.0088	0.0252	−0.0341	0.0307	0.2692	405

This table compares the average teacher characteristics across those receiving any treatment (teacher has an unmonitored day) versus those who experience no treatment (teacher is always monitored in the given Window). Values shown in columns 1 and 2 are deviations from the school mean. In other words, these are average within-school deviations between treated and untreated. None of the differences appear to be statistically significant. Because treatment occurs at the teacher level, all student-teacher combinations are preserved which is why observations are greater than in Table 1.

Table A8
Balance Check for Window 1 across Teachers
(Outcome: Window 1 Unmonitored Days)

	(1)	(2)	(3)	(4)
Teacher Experience	-0.0391	-0.0423	-0.151	-0.160
	(0.159)	(0.162)	(0.279)	(0.282)
Teacher Experience Squared	0.00200	0.00206	0.00591	0.00611
	(0.00557)	(0.00569)	(0.0102)	(0.0103)
Black		0.149	0.300	-0.201
		(0.900)	(1.492)	(1.549)
Hispanic		1.112	-3.369	-3.266
		(2.674)	(5.680)	(6.271)
Asian		-0.746	-2.456	-2.083
		(2.611)	(3.741)	(3.697)
Lagged Evaluation Score			1.519	1.250
			(1.828)	(1.986)
Highly Effective (t-1)				-3.839*
				(1.983)
Minimally Effective (t-1)				-2.255
				(1.955)
School FE	X	X	X	X
Year FE	X	X	X	X
Observations	678	678	400	400
F	0.116	0.106	0.331	0.884
p	0.890	0.991	0.920	0.530

This table checks if treatment is systematically targeted at certain teachers based on observable characteristics. The outcome is number of unmonitored days in Window 1. Errors are not clustered. The only statistically significant coefficient is that on teachers who were ranked as "Highly Effective" in the previous year. These teachers have less unmonitored time in Window 1, which could negatively bias the effect of unmonitored days. The F-Test is a joint hypothesis test of whether all coefficients are zero (not including the fixed-effects). These too appear to fail to reject the null hypothesis.

Table A9
Balance Check for Window 2 across Teachers
(Outcome: Window 2 Unmonitored Days)

	(1)	(2)	(3)	(4)
Teacher Experience	0.0364	0.0458	-0.0438	-0.0553
	(0.0940)	(0.0968)	(0.117)	(0.115)
Teacher Experience Squared	0.000592	0.000339	0.00154	0.00195
	(0.00358)	(0.00362)	(0.00452)	(0.00449)
Black		-0.0597	-0.207	-0.199
		(0.512)	(0.690)	(0.702)
Hispanic		1.746	-1.322	-1.604
		(2.017)	(1.271)	(1.283)
Asian		0.673	-0.214	-0.336
		(1.087)	(0.360)	(0.412)
Lagged Evaluation Score			-0.345	-0.610
			(0.630)	(0.713)
Highly Effective (t-1)				0.589
				(0.689)
Minimally Effective (t-1)				-0.542
				(0.825)
School FE	X	X	X	X
Year FE	X	X	X	X
Observations	678	678	400	400
F	1.450	0.932	0.415	0.447
p	0.235	0.459	0.869	0.892

This table checks if treatment is systematically targeted at certain teachers based on observable characteristics. The outcome is number of unmonitored days in Window 2. None of the coefficients are significant, supporting the identification assumption. The F-Test is a joint hypothesis test of whether all coefficients are zero (not including the fixed-effects). These too appear to fail to reject the null hypothesis.

Table A10
Balance Check for Window 3 across Teachers
(Outcome: Window 3 Unmonitored Days)

	(1)	(2)	(3)	(4)
Teacher Experience	-0.0169	-0.0125	0.0126	0.0118
	(0.0620)	(0.0614)	(0.0951)	(0.0970)
Teacher Experience Squared	0.000998	0.000921	0.000263	0.000297
	(0.00219)	(0.00217)	(0.00345)	(0.00351)
Black		-0.0796	-0.0544	-0.0360
		(0.308)	(0.417)	(0.427)
Hispanic		-0.136	2.873	2.842
		(1.012)	(2.158)	(2.195)
Asian		1.149	0.950	0.925
		(1.127)	(1.081)	(1.099)
Lagged Evaluation Score			-0.316	-0.333
			(0.472)	(0.490)
Highly Effective (t-1)				0.193
				(0.491)
Minimally Effective (t-1)				0.0254
				(0.630)
School FE	X	X	X	X
Year FE	X	X	X	X
Observations	678	678	400	400
F	0.275	0.330	0.588	0.461
p	0.760	0.895	0.740	0.883

This table checks if treatment is systematically targeted at certain teachers based on observable characteristics. The outcome is number of unmonitored days in Window 3. None of the coefficients are significant, supporting the identification assumption. The F-Test is a joint hypothesis test of whether all coefficients are zero (not including the fixed-effects). These too appear to fail to reject the null hypothesis.

Table A11
Balance Check for Space Between Evaluation $M1$ and Feedback across Teachers
(Outcome Measured in Days)

	(1)	(2)	(3)	(4)
Teacher Experience	0.0865	0.127	0.0744	0.100
	(0.125)	(0.120)	(0.124)	(0.128)
Teacher Experience Squared	-0.00288	-0.00400	-0.00236	-0.00327
	(0.00442)	(0.00433)	(0.00446)	(0.00457)
Black		-0.896	-1.240*	-1.157
		(0.664)	(0.716)	(0.717)
Hispanic		2.749	2.614	3.182*
		(1.878)	(1.698)	(1.739)
Asian		-1.933	0.385	0.564
		(2.717)	(1.643)	(1.635)
Lagged Evaluation Score			0.933	1.539
			(0.979)	(1.049)
Highly Effective (t-1)				-0.463
				(1.338)
Minimally Effective (t-1)				1.574**
				(0.736)
School FE	X	X	X	X
Year FE	X	X	X	X
Observations	450	450	400	400
F	0.239	1.184	1.184	1.442
p	0.788	0.317	0.315	0.178

This table checks if the space between evaluation and subsequent feedback for the $M1$ evaluation systematically changes for certain teachers based on observable characteristics. Teachers rated as "Minimally Effective" in the previous year appear to have 1.3 days more time between evaluation and feedback for their $M1$ evaluation, which could positively bias my estimate of the effect of feedback from $M1$. The F-Test is a joint hypothesis test of whether all coefficients are zero (not including the fixed-effects). These appear to fail to reject the null hypothesis.

Table A12
Balance Check for Space Between Evaluation $M2$ and Feedback across Teachers
(Outcome Measured in Days)

	(1)	(2)	(3)	(4)
Teacher Experience	-0.112	-0.165	-0.0592	-0.0591
	(0.118)	(0.120)	(0.130)	(0.134)
Teacher Experience Squared	0.00364	0.00488	0.000854	0.000832
	(0.00450)	(0.00451)	(0.00496)	(0.00509)
Black		1.192*	1.361**	1.298*
		(0.623)	(0.678)	(0.680)
Hispanic		2.224*	1.923	1.968
		(1.152)	(2.115)	(2.254)
Asian		-2.531**	-2.553*	-2.492*
		(1.225)	(1.312)	(1.285)
Lagged Evaluation Score			-2.057**	-2.060*
			(0.915)	(1.048)
Highly Effective (t-1)				-0.544
				(0.849)
Minimally Effective (t-1)				-0.218
				(1.062)
School FE	X	X	X	X
Year FE	X	X	X	X
Observations	450	450	400	400
F	0.511	2.708	2.101	1.702
p	0.600	0.0203	0.0528	0.0972

This table checks if the space between evaluation and subsequent feedback for the $M2$ evaluation systematically changes for certain teachers based on observable characteristics. Teachers with higher previous evaluation scores appear to have received less time between their evaluation and feedback, which would positively bias the effect of $M2$ feedback. The F-Test is a joint hypothesis test of whether all coefficients are zero (not including the fixed-effects). These fail to reject the null hypothesis in two cases, suggesting there are some correlations between teacher characteristics and feedback time.

Table A13
Balance Check for Space Between Evaluation $P1$ and Feedback across Teachers
(Outcome Measured in Days)

	(1)	(2)	(3)	(4)
Teacher Experience	0.199	0.188	0.215	0.205
	(0.191)	(0.185)	(0.202)	(0.206)
Teacher Experience Squared	-0.00646	-0.00636	-0.00770	-0.00741
	(0.00691)	(0.00675)	(0.00745)	(0.00757)
Black		0.155	0.234	-0.0534
		(1.014)	(1.149)	(1.200)
Hispanic		1.521	0.332	0.277
		(1.248)	(1.612)	(1.331)
Asian		-4.925	-5.160	-4.993
		(3.216)	(3.345)	(3.336)
Lagged Evaluation Score			0.541	0.278
			(1.215)	(1.222)
Highly Effective (t-1)				-1.987*
				(1.103)
Minimally Effective (t-1)				-1.524
				(1.253)
School FE	X	X	X	X
Year FE	X	X	X	X
Observations	450	450	400	400
F	0.605	1.321	0.940	1.361
p	0.547	0.254	0.466	0.213

This table checks if the space between evaluation and subsequent feedback for the $P1$ evaluation systematically changes for certain teachers based on observable characteristics. Teachers who were previously rated as "Highly Effective" had fewer days between their $P1$ evaluation and feedback, which would positively bias my estimate for the effect of feedback. The F-Test is a joint hypothesis test of whether all coefficients are zero (not including the fixed-effects). These fail to reject the null hypothesis.

Table A14
Balance Check for Space Between Evaluation $P2$ and Feedback across Teachers
(Outcome Measured in Days)

	(1)	(2)	(3)	(4)
Teacher Experience	0.0763	0.101	0.0601	0.0764
	(0.141)	(0.145)	(0.165)	(0.165)
Teacher Experience Squared	-0.00226	-0.00278	-0.00107	-0.00161
	(0.00520)	(0.00527)	(0.00600)	(0.00601)
Black		-0.505	-0.244	-0.0934
		(0.647)	(0.725)	(0.742)
Hispanic		-0.890	0.352	0.647
		(1.142)	(1.268)	(1.512)
Asian		2.830	3.111	3.133
		(3.287)	(3.395)	(3.464)
Lagged Evaluation Score			0.566	0.959
			(0.823)	(0.865)
Highly Effective (t-1)				0.538
				(1.189)
Minimally Effective (t-1)				1.345
				(0.873)
School FE	X	X	X	X
Year FE	X	X	X	X
Observations	450	450	400	400
F	0.204	0.433	0.314	0.635
p	0.815	0.825	0.929	0.748

This table checks if the space between evaluation and subsequent feedback for the $P2$ evaluation systematically changes for certain teachers based on observable characteristics. Teachers who were previously rated as "Minimally Effective" had more days between their $P2$ evaluation and feedback, which would positively bias my estimate for the effect of feedback. The F-Test is a joint hypothesis test of whether all coefficients are zero (not including the fixed-effects). These fail to reject the null hypothesis.

Complete Versions of Tables 10 and 11

Table A15
Effect of Unmonitored Time on Student Math Test Outcomes
(Outcome: Math Standard Deviations)

	(1)	(2)	(3)	(4)	(5)
Window 1	0.0013	0.0011	0.0018	0.0014	0.0009
	(0.0013)	(0.0020)	(0.0022)	(0.0021)	(0.0021)
	[0.2042 0.2574]	[0.5628 0.6246]	[0.5276 0.5901]	[0.6175 0.6776]	[0.7541 0.8063]
Window 2	0.0018	-0.0058	-0.0054	-0.0072	-0.0092
	(0.0033)	(0.0074)	(0.0074)	(0.0072)	(0.0071)
	[0.5861 0.6472]	[0.1014 0.1428]	[0.1219 0.1662]	[0.0356 0.0631]	[0.0055 0.0196]
Window 3	-0.0105	-0.0130	-0.0134	-0.0127	-0.0112
	(0.0040)	(0.0067)	(0.0066)	(0.0063)	(0.0063)
	[0.0461 0.0766]	[0.0123 0.0307]	[0.0084 0.0246]	[0.0154 0.0355]	[0.0270 0.0518]
P1 Post-Feedback (Days)		0.0015	0.0011	0.0011	0.0012
		(0.0015)	(0.0016)	(0.0015)	(0.0015)
P2 Post-Feedback (Days)		0.0047	0.0040	0.0051	0.0060
		(0.0045)	(0.0047)	(0.0045)	(0.0045)
M1 Post-Feedback (Days)		-0.0003	0.0003	0.0005	0.0004
		(0.0008)	(0.0009)	(0.0009)	(0.0009)
M2 Post-Feedback (Days)		0.0009	0.0010	0.0007	0.0009
		(0.0023)	(0.0024)	(0.0024)	(0.0024)
P2 Cumulative Monitoring Intensity			-0.0199	-0.0181	-0.0215
			(0.0259)	(0.0253)	(0.0244)
M1 Cumulative Monitoring Intensity			0.0367	0.0323	0.0295
			(0.0383)	(0.0377)	(0.0366)
M2 Cumulative Monitoring Intensity			0.0089	0.0066	0.0134
			(0.0402)	(0.0402)	(0.0401)
Previous Student Scores	X	X	X	X	X
Teacher Experience				X	X
Student Demographics					X
Observations	12305	12305	12305	12305	12305

This table demonstrates the key results for student math outcomes. All standard errors are clustered at the teacher-year level. The 95% confidence intervals for p-values from 1,000 Randomization Inference trials are shown in brackets. Sample includes students with a previous year's test score and a teacher with one or more years of experience.

Table A16
Effect of Unmonitored Time on Student Reading Test Outcomes
(Outcome: Reading Standard Deviations)

	(1)	(2)	(3)	(4)	(5)
Window 1	0.0010	0.0037	0.0039	0.0040	0.0038
	(0.0011)	(0.0019)	(0.0019)	(0.0019)	(0.0018)
	[0.3478 0.4089]	[0.0162 0.0367]	[0.3714 0.4331]	[0.3675 0.4291]	[0.3616 0.4230]
Window 2	0.0042	0.0051	0.0049	0.0052	0.0027
	(0.0023)	(0.0053)	(0.0052)	(0.0051)	(0.0052)
	[0.0959 0.1364]	[0.0391 0.0676]	[0.0399 0.0688]	[0.0356 0.0631]	[0.2553 0.3120]
Window 3	-0.0117	-0.0144	-0.0137	-0.0142	-0.0122
	(0.0030)	(0.0052)	(0.0051)	(0.0053)	(0.0052)
	[0.0028 0.0144]	[0.0000 0.0056]	[0.0000 0.0037]	[0.0000 0.0056]	[0.0011 0.0102]
P1 Post-Feedback (Days)		0.0006	0.0006	0.0005	0.0005
		(0.0012)	(0.0012)	(0.0012)	(0.0012)
P2 Post-Feedback (Days)		-0.0009	-0.0018	-0.0019	-0.0008
		(0.0034)	(0.0034)	(0.0034)	(0.0034)
M1 Post-Feedback (Days)		-0.0018	-0.0018	-0.0018	-0.0018
		(0.0007)	(0.0008)	(0.0008)	(0.0007)
M2 Post-Feedback (Days)		0.0011	-0.0010	-0.0009	-0.0005
		(0.0016)	(0.0019)	(0.0020)	(0.0019)
P2 Cumulative Monitoring Intensity			-0.0253	-0.0249	-0.0200
			(0.0176)	(0.0178)	(0.0171)
M1 Cumulative Monitoring Intensity			0.0036	0.0042	0.0080
			(0.0292)	(0.0289)	(0.0270)
M2 Cumulative Monitoring Intensity			-0.0835	-0.0863	-0.0679
			(0.0384)	(0.0386)	(0.0361)
Previous Student Scores	X	X	X	X	X
Teacher Experience				X	X
Student Demographics					X
Observations	12820	12820	12820	12820	12820

This table demonstrates the key results for student reading outcomes. All standard errors are clustered at the teacher-year level. The 95% confidence intervals for p-values from 1,000 Randomization Inference trials are shown in brackets. Sample includes students with a previous year's test score and a teacher with one or more years of experience.

Different Specifications of Experience

Table A17
Math Results under Different Specifications of Teacher Experience
(Outcome: Math Standard Deviations)

	(1)	(2)	(3)	(4)
Window 1	0.0009	0.0001	0.0008	-0.0014
	(0.0021)	(0.0021)	(0.0021)	(0.0020)
Window 2	-0.0092	-0.0092	-0.0085	-0.0071
	(0.0071)	(0.0069)	(0.0073)	(0.0069)
Window 3	-0.0112	-0.0114	-0.0112	-0.0081
	(0.0063)	(0.0066)	(0.0062)	(0.0064)
Previous Student Scores	X	X	X	X
Post-evaluation Feedback Time	X	X	X	X
Evaluation Probability Controls	X	X	X	X
Student Race and Gender	X	X	X	X
Other Student Demographics	X	X	X	X
Observations	12305	12305	12305	12305

This table demonstrates the key results for student math outcomes where each column uses a different measurement of teacher experience. All standard errors are clustered at the teacher-year level. Column (1): Experience is a continuous variable capped at 20 years; both linear and squared terms included. Column (2): Experience is a categorical variable capped at 20 years. Column (3): Same as Column (1) but not capped at 20 years. Column (4): Same as Column (2) but not capped at 20 years.

Table A18
Reading Results under Different Specifications of Teacher Experience
(Outcome: Reading Standard Deviations)

	(1)	(2)	(3)	(4)
Window 1	0.0038	0.0031	0.0037	0.0026
	(0.0018)	(0.0019)	(0.0019)	(0.0020)
Window 2	0.0027	0.0055	0.0017	0.0055
	(0.0052)	(0.0055)	(0.0053)	(0.0057)
Window 3	-0.0122	-0.0122	-0.0116	-0.0154
	(0.0052)	(0.0056)	(0.0051)	(0.0065)
Previous Student Scores	X	X	X	X
Post-evaluation Feedback Time	X	X	X	X
Evaluation Probability Controls	X	X	X	X
Student Race and Gender	X	X	X	X
Other Student Demographics	X	X	X	X
Observations	12820	12820	12820	12820

This table demonstrates the key results for student reading outcomes where each column uses a different measurement of teacher experience. All standard errors are clustered at the teacher-year level. Column (1): Experience is a continuous variable capped at 20 years; both linear and squared terms included. Column (2): Experience is a categorical variable capped at 20 years. Column (3): Same as Column (1) but not capped at 20 years. Column (4): Same as Column (2) but not capped at 20 years.

Excluding Students with Suspensions

Table A19
Math Results Without Students with Any Suspensions
(Outcome: Math Standard Deviations)

	(1)	(2)	(3)
Window 1	0.0002	0.0004	0.0002
	(0.0012)	(0.0012)	(0.0012)
	[0.9741 0.9908]	[0.9047 0.9388]	[0.9482 0.9730]
Window 2	-0.0065	-0.0063	-0.0062
	(0.0039)	(0.0040)	(0.0040)
	[0.2814 0.3397]	[0.3058 0.3652]	[0.2989 0.3581]
Window 3	-0.0089	-0.0093	-0.0081
	(0.0053)	(0.0051)	(0.0050)
	[0.1995 0.2522]	[0.1861 0.2376]	[0.2398 0.2956]
Previous Student Scores	X	X	X
Teacher Experience		X	X
Student Demographics			X
Observations	7416	7416	7416

This table demonstrates the key results for student math outcomes. All students that have received any suspensions in the year are dropped. All standard errors are clustered at the teacher-year level. The 95% confidence intervals for p-values from 1,000 Randomization Inference trials are shown in brackets. The results mirror Table 10. The point estimates on Window 3 are similar but vary more between specifications. Estimates using feedback time and evaluation probability are excluded due to the greatly reduced sample size.

Table A20
Reading Results Without Students with Any Suspensions
(Outcome: Reading Standard Deviations)

	(1)	(2)	(3)
Window 1	0.0012	0.0012	0.0010
	(0.0008)	(0.0009)	(0.0008)
	[0.6736 0.7312]	[0.6664 0.7244]	[0.7147 0.7698]
Window 2	-0.0043	-0.0046	-0.0054
	(0.0025)	(0.0024)	(0.0028)
	[0.2669 0.3243]	[0.2485 0.3048]	[0.1510 0.1989]
Window 3	-0.0064	-0.0068	-0.0051
	(0.0030)	(0.0032)	(0.0030)
	[0.1937 0.2459]	[0.1832 0.2345]	[0.2912 0.3499]
Previous Student Scores	X	X	X
Teacher Experience		X	X
Student Demographics			X
Observations	7787	7787	7787

This table demonstrates the key results for student reading outcomes. All students that have received any suspensions in the year are dropped. All standard errors are clustered at the teacher-year level. The 95% confidence intervals for p-values from 1,000 Randomization Inference trials are shown in brackets. The results mirror Table 10. The point estimates on Window 3 are similar but vary more between specifications. Estimates using feedback time and evaluation probability are excluded due to the greatly reduced sample size.

Different Specifications of Expected Evaluation Distribution

Table A21
Math Results Assuming Uniform Evaluation Distribution Expectation
(Outcome: Math Standard Deviations)

	(1)	(2)	(3)	(4)	(5)
Window 1	0.0013	0.0011	0.0011	0.0006	0.0000
	(0.0013)	(0.0020)	(0.0021)	(0.0021)	(0.0020)
Window 2	0.0018	-0.0058	-0.0055	-0.0075	-0.0096
	(0.0033)	(0.0074)	(0.0074)	(0.0072)	(0.0072)
Window 3	-0.0105	-0.0130	-0.0137	-0.0129	-0.0114
	(0.0040)	(0.0067)	(0.0065)	(0.0062)	(0.0062)
Previous Student Scores	X	X	X	X	X
Post-evaluation Feedback Time		X	X	X	X
Evaluation Probability Controls			X	X	X
Teacher Experience				X	X
Student Demographics					X
Observations	12305	12305	12305	12305	12305

This table demonstrates the key results for student math outcomes. Columns 1 through 5 are the same as in Table 10 but are included for comparison purposes. All standard errors are clustered at the teacher-year level. The results for Window 3 unmonitored time mirror Table 10 identically, reinforcing the notion that (a) the method for measuring the probability of an evaluation does not matter for this specification, and (b) teachers' behavioral changes as the result of evaluation probability do not have a meaningful effect on student outcomes.

Table A22
Reading Results Assuming Uniform Evaluation Distribution Expectation
(Outcome: Reading Standard Deviations)

	(1)	(2)	(3)	(4)	(5)
Window 1	0.0010	0.0037	0.0030	0.0031	0.0029
	(0.0011)	(0.0019)	(0.0019)	(0.0019)	(0.0019)
Window 2	0.0042	0.0051	0.0050	0.0053	0.0029
	(0.0023)	(0.0053)	(0.0050)	(0.0049)	(0.0050)
Window 3	-0.0117	-0.0144	-0.0140	-0.0143	-0.0123
	(0.0030)	(0.0052)	(0.0052)	(0.0054)	(0.0053)
Previous Student Scores	X	X	X	X	X
Post-evaluation Feedback Time		X	X	X	X
Evaluation Probability Controls			X	X	X
Teacher Experience				X	X
Student Demographics					X
Observations	12820	12820	12820	12820	12820

This table demonstrates the key results for student reading outcomes. Columns 1 through 5 are the same as in Table 11 but are included for comparison purposes. All standard errors are clustered at the teacher-year level. The results for Window 3 unmonitored time mirror Table 10 closely, reinforcing the notion that (a) the method for measuring the probability of an evaluation does not matter for this specification, and (b) teachers' behavioral changes as the result of evaluation probability do not have a meaningful effect on student outcomes.

Heterogeneous Results by Grade

Table A23
Math Results by Grade
(Outcome: Math Standard Deviations)

	(1)	(2)	(3)	(4)	(5)
4th Grade × Window 1	0.0030	0.0027	0.0034	0.0030	0.0026
	(0.0014)	(0.0018)	(0.0021)	(0.0021)	(0.0020)
5th Grade × Window 1	0.0001	-0.0004	0.0002	-0.0002	-0.0007
	(0.0014)	(0.0020)	(0.0022)	(0.0022)	(0.0021)
4th Grade × Window 2	-0.0029	-0.0133	-0.0127	-0.0137	-0.0139
	(0.0046)	(0.0085)	(0.0085)	(0.0083)	(0.0082)
5th Grade × Window 2	0.0054	-0.0037	-0.0029	-0.0047	-0.0076
	(0.0049)	(0.0078)	(0.0077)	(0.0075)	(0.0074)
4th Grade × Window 3	-0.0177	-0.0207	-0.0205	-0.0194	-0.0177
	(0.0047)	(0.0071)	(0.0072)	(0.0071)	(0.0070)
5th Grade × Window 3	-0.0063	-0.0092	-0.0099	-0.0096	-0.0082
	(0.0052)	(0.0069)	(0.0068)	(0.0064)	(0.0064)
Previous Student Scores	X	X	X	X	X
Post-evaluation Feedback Time		X	X	X	X
Evaluation Probability Controls			X	X	X
Teacher Experience				X	X
Student Demographics					X
Observations	12305	12305	12305	12305	12305

This table demonstrates the key results for student math outcomes broken out by grade. All errors are clustered at the teacher-year level. Sample includes students with a previous year's test score and a teacher with one or more years of experience. Effects of unmonitored time in Window 3 are more concentrated in fourth grade than in fifth. The fourth grade math curriculum covers fractions and operations on fractions, while the fifth grade math curriculum covers decimals and their operations.

Table A24
Reading Results by Grade
(Outcome: Reading Standard Deviations)

	(1)	(2)	(3)	(4)	(5)
4th Grade × Window 1	0.0018	0.0043	0.0046	0.0046	0.0047
	(0.0012)	(0.0018)	(0.0019)	(0.0019)	(0.0018)
5th Grade × Window 1	0.0004	0.0028	0.0029	0.0031	0.0027
	(0.0012)	(0.0020)	(0.0020)	(0.0020)	(0.0019)
4th Grade × Window 2	0.0014	0.0022	0.0020	0.0020	0.0021
	(0.0033)	(0.0063)	(0.0061)	(0.0060)	(0.0060)
5th Grade × Window 2	0.0064	0.0069	0.0068	0.0071	0.0030
	(0.0032)	(0.0057)	(0.0055)	(0.0054)	(0.0054)
4th Grade × Window 3	-0.0108	-0.0133	-0.0128	-0.0131	-0.0133
	(0.0056)	(0.0070)	(0.0069)	(0.0072)	(0.0070)
5th Grade × Window 3	-0.0120	-0.0138	-0.0129	-0.0134	-0.0110
	(0.0030)	(0.0050)	(0.0048)	(0.0050)	(0.0049)
Previous Student Scores	X	X	X	X	X
Post-evaluation Feedback Time		X	X	X	X
Evaluation Probability Controls			X	X	X
Teacher Experience				X	X
Student Demographics					X
Observations	12820	12820	12820	12820	12820

This table demonstrates the key results for student reading outcomes broken out by grade. All errors are clustered at the teacher-year level. Sample includes students with a previous year's test score and a teacher with one or more years of experience.

Results Including Inexperienced Teachers

Table A25
Math Results Including Inexperienced Teachers
(Outcome: Math Standard Deviations)

	(1)	(2)	(3)	(4)	(5)
Veteran × Window 1	0.0006	0.0001	0.0014	0.0011	0.0006
	(0.0013)	(0.0014)	(0.0020)	(0.0020)	(0.0019)
First-Year × Window 1	-0.0016	0.0036	0.0051	0.0049	0.0047
	(0.0061)	(0.0059)	(0.0062)	(0.0062)	(0.0060)
Veteran × Window 2	0.0018	-0.0058	-0.0049	-0.0065	-0.0085
	(0.0033)	(0.0073)	(0.0073)	(0.0071)	(0.0070)
First-Year × Window 2	0.0067	-0.0050	-0.0062	-0.0079	-0.0066
	(0.0099)	(0.0194)	(0.0191)	(0.0190)	(0.0189)
Veteran × Window 3	-0.0093	-0.0119	-0.0121	-0.0114	-0.0099
	(0.0040)	(0.0067)	(0.0066)	(0.0063)	(0.0063)
First-Year × Window 3	0.0673	-0.0569	-0.0576	-0.0577	-0.0530
	(0.0183)	(0.0351)	(0.0351)	(0.0351)	(0.0343)
Previous Student Scores	X	X	X	X	X
Post-evaluation Feedback Time		X	X	X	X
Evaluation Probability Controls			X	X	X
Teacher Experience				X	X
Student Demographics					X
Observations	13080	13080	13080	13080	13080

This table demonstrates the key results for student reading outcomes broken out by whether or not the teacher is in her first year of teaching. All errors are clustered at the teacher-year level. The sample is all students with a previous year's test score. Unmonitored time appears to have a positive effect for students of first-year teachers. However, once controlling for feedback, the results reverse. Though not shown, the coefficients for post-feedback effects are positive but insignificant for first-year teachers, as in the main results. First-year teachers appear to have markedly different responses both to unmonitored time and feedback, but the small sample size limits the accuracy of these measurements.

Table A26
Reading Results Including Inexperienced Teachers
(Outcome: Reading Standard Deviations)

	(1)	(2)	(3)	(4)	(5)
Veteran × Window 1	0.0009	0.0000	0.0016	0.0017	0.0015
	(0.0011)	(0.0012)	(0.0017)	(0.0017)	(0.0016)
First-Year × Window 1	0.0071	0.0112	0.0132	0.0132	0.0129
	(0.0035)	(0.0038)	(0.0040)	(0.0040)	(0.0042)
Veteran × Window 2	0.0042	0.0053	0.0056	0.0057	0.0031
	(0.0023)	(0.0054)	(0.0053)	(0.0052)	(0.0053)
First-Year × Window 2	0.0165	0.0041	0.0026	0.0033	-0.0008
	(0.0055)	(0.0140)	(0.0135)	(0.0136)	(0.0127)
Veteran × Window 3	-0.0115	-0.0146	-0.0145	-0.0149	-0.0127
	(0.0030)	(0.0053)	(0.0052)	(0.0054)	(0.0053)
First-Year × Window 3	0.0522	-0.0286	-0.0322	-0.0318	-0.0379
	(0.0134)	(0.0357)	(0.0349)	(0.0349)	(0.0387)
Previous Student Scores	X	X	X	X	X
Post-evaluation Feedback Time		X	X	X	X
Evaluation Probability Controls			X	X	X
Teacher Experience				X	X
Student Demographics					X
Observations	13669	13667	13669	13669	13669

This table demonstrates the key results for student reading outcomes broken out by whether or not the teacher is in her first year of teaching. All errors are clustered at the teacher-year level. The sample is all students with a previous year's test score. Unmonitored time appears to have a positive effect for students of first-year teachers. However, once controlling for feedback, the results reverse. Though not shown, the coefficients for post-feedback effects are positive but insignificant for first-year teachers, as in the main results. First-year teachers appear to have markedly different responses both to unmonitored time and feedback, but the small sample size limits the accuracy of these measurements.

Table A27
Description of the Components of the Teaching and Learning Framework

STANDARD	DESCRIPTION OF HIGHLY EFFECTIVE TEACHING
Teach 1 *Lead well-organized, objective-driven lessons*	*Lesson Organization* The lesson is well-organized: All parts of the lesson are connected to each other and aligned to the objective, and each part significantly moves all students toward mastery of the objective. *Lesson Objective* The objective of the lesson is clear to students and conveys what students are learning and what they will be able to do as a result of the lesson. Students also can authentically explain what they are learning and doing beyond simply repeating the stated or posted objective. *Objective Importance* Students understand the importance of the objective. Students also can authentically explain why what they are learning and doing is important, beyond simply repeating the teachers' explanation.
Teach 2 *Explain content clearly*	*Clear, Coherent Delivery* Explanations of content are clear and coherent, and they build student understanding of content. The teacher might provide explanations through direct verbal or written delivery, modeling or demonstrations, think-alouds, visuals, or questioning. Explanations of content also are delivered in as direct and efficient a manner as possible. *Academic Language* The teacher gives clear, precise definitions and uses a broad vocabulary that includes specific academic language and words that may be unfamiliar to students when it is appropriate to do so. Students also demonstrate through their verbal or written responses that they are internalizing academic vocabulary. *Emphasize Key Points* The teacher emphasizes key points when necessary, such that students understand the main ideas of the content. Students also can authentically explain the main ideas of the content beyond simply repeating back the teacher's explanations. *Student Understanding* Students show that they understand the explanations. When appropriate, concepts also are explained in a way that actively and effectively involves students in the learning process. For example, students have opportunities to explain concepts to each other. *Connections* The teacher makes connections with students' prior knowledge, students' experiences and interests, other content areas, or current events to effectively build student understanding of content.

Table A27
(Continued)

STANDARD	DESCRIPTION OF HIGHLY EFFECTIVE TEACHING
Teach 3 *Engage students at all learning levels in accessible and challenging work*	*Accessibility* The teacher makes the lesson accessible to all students. There is evidence that the teacher knows each student's level and ensures that the lesson meets all students where they are. *Challenge* The teacher makes the lesson challenging to all students. There is evidence that the teacher knows each student's level and ensures that the lesson pushes all students forward from where they are. *Balance* There is an appropriate balance between teacher-directed and student-centered learning during the lesson, such that students have adequate opportunities to meaningfully practice, apply, and demonstrate what they are learning.
Teach 4 *Provide students multiple ways to move toward mastery*	*Multiple Ways Toward Mastery* The teacher provides students multiple ways to engage with content, and all ways move students toward mastery of lesson content. During the lesson, students are also developing deep understanding of the content. *Appropriateness for Students* The ways the teacher provides include learning styles or modalities that are appropriate to students' needs; all students respond positively and are actively involved in the work.
Teach 5 *Check for student understanding*	*Key Moments* The teacher checks for understanding of content at all key moments. *Accurate Pulse* The teacher always gets an accurate "pulse" at key moments by using one or more checks that gather information about the depth of understanding for a range of students, when appropriate.

Table A27
(Continued)

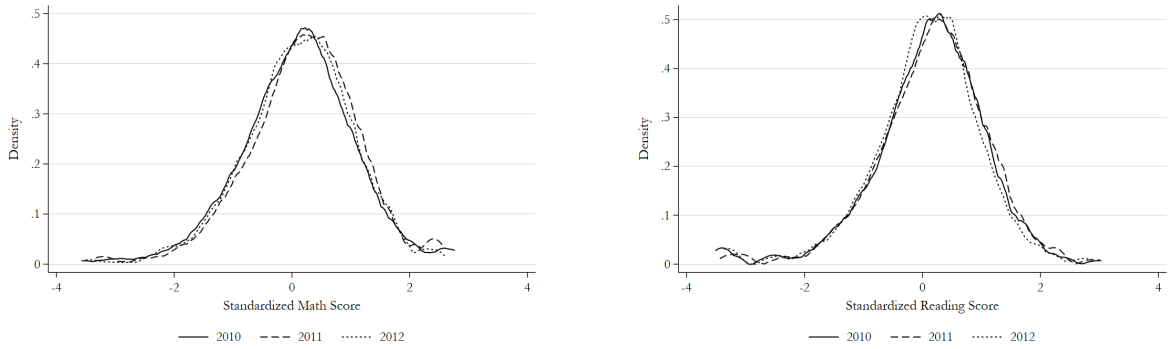
STANDARD	DESCRIPTION OF HIGHLY EFFECTIVE TEACHING
Teach 6 *Respond to student understanding*	*Scaffolding* When students demonstrate misunderstandings or partial understandings, the teacher always uses effective scaffolding techniques that enable students to construct their own understandings, when appropriate. *Re-Teaching* The teacher always re-teaches effectively when appropriate, such as in cases in which most of the class demonstrates a misunderstanding or an individual student demonstrates a significant misunderstanding. The teacher also anticipates common misunderstandings (e.g., by offering a misunderstanding as a correct answer to see how students respond) or recognizes a student response as a common misunderstanding and shares it with the class to lead all students to a more complete understanding. *Probing* The teacher always probes students' correct responses, when appropriate, to ensure student understanding.
Teach 7 *Develop higher-level understanding through effective questioning*	*Questions and Tasks* The teacher asks questions that push all students' thinking; when appropriate, the teacher also poses tasks that are increasingly complex that develop all students' higher-level understanding. *Support* After posing a question or task, the teacher always uses appropriate strategies to ensure that students move toward higher-level understanding. *Meaningful Response* Almost all students answer questions of complete complex tasks with meaningful responses that demonstrate movement toward higher-level understanding, showing that they are accustomed to being asked these kinds of questions.

Table A27
(Continued)

STANDARD	DESCRIPTION OF HIGHLY EFFECTIVE TEACHING
Teach 8 *Maximize instructional time*	*Routines, Procedures, and Transitions* Routines, procedures, and transitions are orderly, efficient, and systematic with minimal prompting from the teacher' students know their responsibilities and some students share responsibility for leading the operations and routines in the classroom. *Student Idleness* Students always have something meaningful to do. Lesson pacing is also student-directed or individualized, when appropriate. *Lesson Pacing* The teacher spends an appropriate amount of time on each part of the lesson. *Student Behavior* Inappropriate or off-task student behavior never interrupts or delays the lesson, either because no such behavior occurs or because when such behavior occurs the teacher efficiently addresses it.
Teach 9 *Build a supportive, learning-focused classroom community*	*Investment* Students are invested in their work and value academic success. Students are also invested in the success of their peers. For example, students can be seen helping each other or showing interest in other students' work without prompting from the teacher. *Risk-Taking* The classroom environment is safe for students, such that students are willing to take on challenges and risk failure. For example, students are eager to ask questions, feel comfortable asking the teacher for help, feel comfortable engaging in constructive feedback with their classmates, and do not respond negatively when a peer answers a question incorrectly. *Respect* Students are always respectful of the teacher and their peers. For example, students listen and do not interrupt when their peers ask or answer questions. *Reinforcement* The teacher meaningfully reinforces positive behavior and good academic work, when appropriate. Students also give unsolicited praise or encouragement to their peers, when appropriate. *Rapport* The teacher has a positive rapport with students, as demonstrated by displays of positive affect, evidence of relationship building, and expressions of interest in students' thoughts and opinions. There is also evidence that the teacher has strong, individualized relationships with some students in the class.

Additional Figures

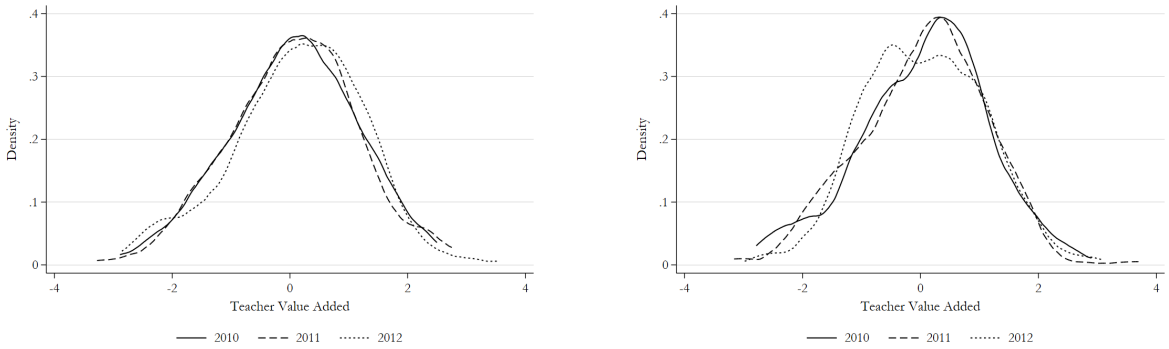
Figure A1
Distribution of Student Math and Reading Scores



(a) Math (b) Reading

Note: These calculations are conducted using only students in the sample. As usual, student standardized test scores are mean-centered for each year.

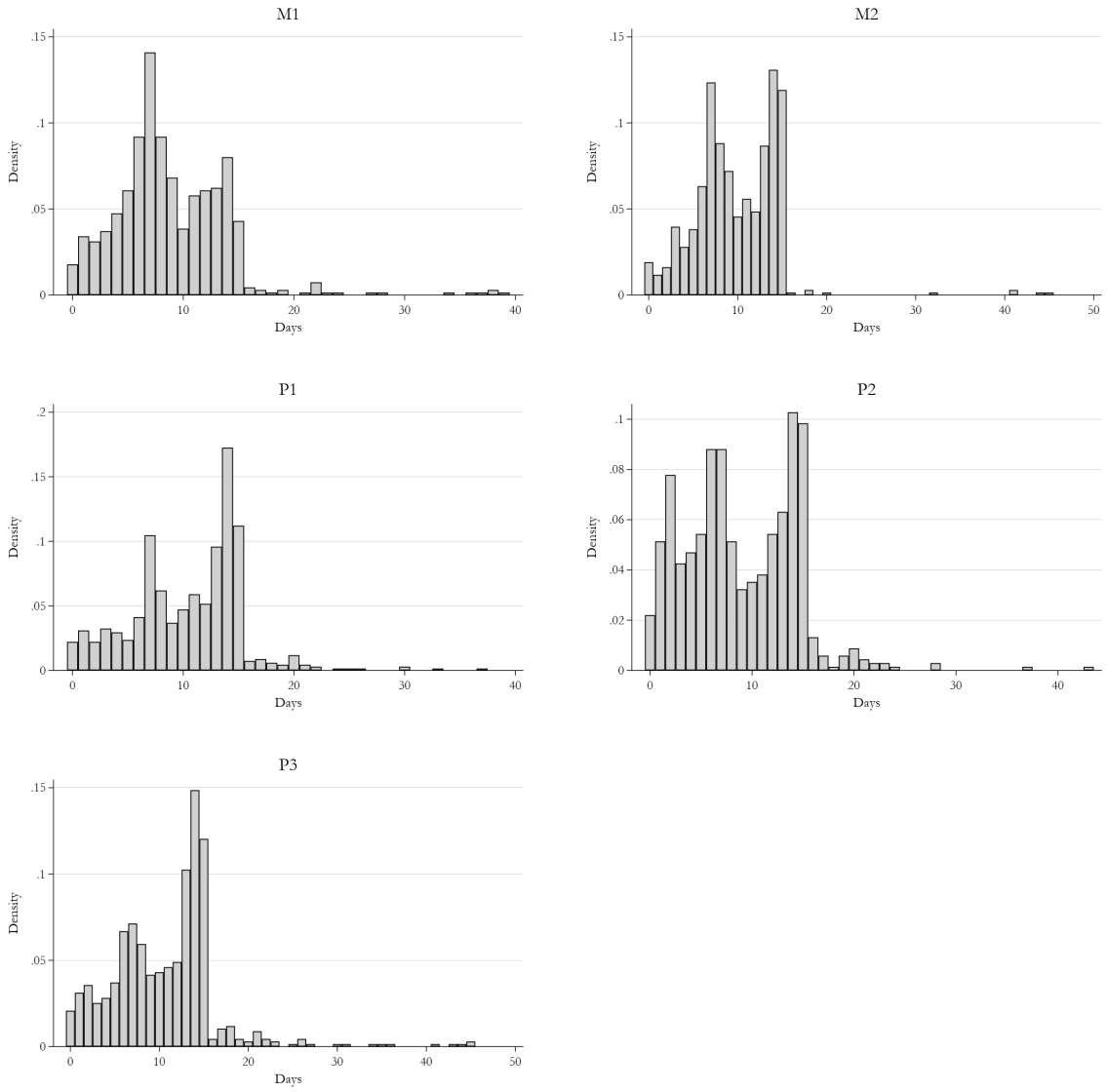
Figure A2
Distribution of Teacher Value-added Scores in Math and Reading



(a) Math (b) Reading

Note: These densities are calculated using only teachers in the sample. Value-added scores are as reported by DCPS, not the author's own calculations.

Figure A3
Histograms of Space between Evaluation and Feedback Conference



Note: Evaluators were expected to provide feedback within three weeks of completing an observation, which explains the sharp dropoff after 15 business days. Some exceptional circumstances prolonged the time.